National Library
of Canada

Canadian Theses Service

Ottawa, Canada
K1A 0N4

Bibliothèque nationale
du Canada

Services des thèses canadiennes

CANADIAN THESES

THÈSES CANADIENNES

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

NL-339 (r. 86/01)

Canada

161

National Library
of Canada

Canadian Theses Division

Ottawa, Canada
K1A 0N4

Bibliothèque nationale
du Canada

Division des thèses canadiennes

0-315-26893-X

PERMISSION TO MICROFILM — AUTORISATION DE MICROFILMER

- Please print or type — Écrire en lettres moulées ou dactylographier

Full Name of Author — Nom complet de l'auteur

DAVID WESLEY MATHESON

Date of Birth — Date de naissance

AUG 29, 1956

Country of Birth — Lieu de naissance

CANADA

Permanent Address — Résidence fixe

#206 - 10175 113 STREET
EDMONTON ALBERTA

Title of Thesis — Titre de la thèse

CONFIRMATORY FACTOR ANALYSIS OF THE WISC-R FOR INUIT CHILDREN

University — Université

UNIVERSITY OF ALBERTA

Degree for which thesis was presented — Grade pour lequel cette thèse fut présentée

MASTER OF EDUCATION

Year this degree conferred — Année d'obtention de ce grade

1983

Name of Supervisor — Nom du directeur de thèse

Dr. R.F. MULCAHY

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

Date

Oct 12, 1983

Signature

David Matheson

THE UNIVERSITY OF ALBERTA

Confirmatory Factor Analysis of the WISC-R for Inuit Children

by

© David Wesley Matheson

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE
OF Master of Education

Department of Educational Psychology

EDMONTON, ALBERTA
FALL, 1983

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR David Wesley Matheson

TITLE OF THESIS Confirmatory Factor Analysis of the WISC-R for Inuit Children

DEGREE FOR WHICH THESIS WAS PRESENTED Master of Education

YEAR THIS DEGREE GRANTED FALL, 1983

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts [illegible]m it may be printed or otherwise reproduced without the author's written permission.

(SIGNED) David Matheson

PERMANENT ADDRESS:

#206 - 10175 - 113 St.
Edmonton, alberta

DATED Oct. 12 1983

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled Confirmatory Factor Analysis of the WISC-R for Inuit Children submitted by David Wesley Matheson in partial fulfilment of the requirements for the degree of Master of Education.

....R Mulcahy....
Supervisor

C Urion....................

T Maguire....................

Date..Sept. 30, 1983..........

Dedication

To the children
of the Keewatin and Kitikmeot Districts,
Northwest Territories

The social scientist is trained to think that he does not know all the answers. The social scientist is not trained to realize that he does not know all the questions. And that is why his social influence is not unfailingly constructive.

Lee J. Cronbach, 1975.

Abstract

The current study assessed the validity of three factor models for WISC-R profile interpretation for use with children in the Keewatin and Kitikmeot districts, Northwest Territories. The models tested were represented by the Verbal and Performance scales; Kaufman's Verbal Comprehension, Perceptual Organization, and Freedom from Distractibility factors; Bannatyne's Conceptualization, Spatial, Sequencing, and Acquired Knowledge factors. The data was comprised of scaled WISC-R subtest scores, which had been collected for standardization purposes from samples of children aged 7 years to 14 yrs, who were enrolled in schools in those districts.

The covariance matrices for the 7 year and 8 year age groups were pooled to increase the sample size for analysis, as were the covariance matrices for each of ages 9 and 10 years, 11 and 12 years, and 13 and 14 years. All eight covariance matrices were also pooled to derive a total sample covariance matrix. Correlation matrices were derived and subjected to principal component, principal factor, and maximum likelihood factor analysis. The latter method allowed direct tests of the fit of each of the three models to the obtained correlation matrices.

The Verbal-Performance model was rejected for all samples, although two factors were sufficient for the 7-8 year and 9-10 year age pools. The Kaufman model provided an acceptable fit to the data for the 13-14 year age pool, whereas slight modifications to that model were required for the 11-12 year age pool. The four-factor Bannatyne model was rejected for every age pool, although the Conceptualization, Spatial, and Sequencing factors were intact within the three-factor solutions for the two older age pools.

The WISC-R is not recommended for clinical use with children in the Keewatin and Kitikmeot districts who are younger than 11 years of age. The use of Kaufman's interpretive model for older children requires additional research to determine the nature of the factors replicated. Various hypotheses concerning unexpected loadings were discussed.

Acknowledgements

I wish to thank my thesis supervisor, Dr. Robert Mulcahy, for his permission to use the WISC-R data collected for the NWT norming project. His encouragement and support for extensive analysis of that data, and his openness to critical appraisal of the validity of the test for NWT children, made the scope of this project feasible.

The other members of the committee, Drs. Tom Maguire and Carl Urion, also offered encouragement, constructive commentary, and numerous hours in digesting various drafts of the thesis. Their ideas and effort were greatly appreciated.

As the Coordinator of Special and Remedial Education for the Government of the Northwest Territories, Ms. Bronwyn Watters was directly responsible for the NWT norming project and coauthored the report of the project with Dr. Mulcahy. Her support for the analysis conducted for this thesis was appreciated, as was her sharing of insights gained from her experience as a psychologist in the Northwest Territories.

Dr. Russ MacArthur of the Department of Educational Psychology contributed to my early awareness and clarification of issues in cross-cultural psychology. His insights and experience are also reflected in this thesis.

I have also benefitted from the countless hours of coffee-saturated discussion, intellectual and otherwise, with fellow students and other faculty members.

Table of Contents

Chapter Page

I. Introduction 1

A. Setting for the Research 1

B. Value and Assumptions of the Present Study 3

C. Advance Organizer to the Literature Review 4

II. Review of Literature 9

A. The Importance of Construct Validity to Assessment 9

Assessment of Intelligence: Purposes and Assumptions 9

Definition and Measurement of Construct Validity 15

Factor Theories of Intelligence 19

Test Bias: Definitions and Significance 34

Summary and Conclusions 43

B. Interpretation of the WISC-R 45

Organization of the Scales 46

Profile Analysis and Interpretive Models 49

Factor Analyses of the WISC-R 52

Derivation and Validity of WISC-R Factor Scores 96

Bias of the WISC-R 123

C. Psychoeducational Assessment of Native Children 129

The WISC-R: Score Patterns and Validity Indices 131

Other Cognitive and Perceptual Tests 140

Inferences on Cognitive Processing 151

Bilingualism and Psychological Assessment ..165

Other Factors Affecting Assessment Results .177

Attempts to Reduce Bias in Assessment 187

Summary 189

D. Testing Factor Models 192

Reliability and Validity of Factor Analysis Results 193

Factorial Invariance and Confirmatory Factor Analysis 198

Maximum Likelihood Methods 210

E. The Present Study 239

III. Methodology 247

A. Subjects 247

Description of Sample 247

Description of Communities 247

A Cautionary Note 258

B. WISC-R 259

C. Data Collection Procedure 264

Testers 264

Test Administration 264

D. Statistical Analysis Procedure 266

Commercial Statistical Software 267

Statistical Notation 269

The LISREL Model 270

Deriving and Testing the Correlation Matrices 274

Factor Analysis Sequence 278

IV. Results 287

A. Tests of Assumptions Regarding the Data 287

Normal Distribution of Scaled Scores 287

Equality of Covariance and Correlation Matrices 291

B. Factor Analysis Results 295

Total Sample 295

Age Pool 7-8 Years 309

Age Pool 9-10 Years 316

Age Pool 11-12 Years 324

Age Pool 13-14 Years 332

Cross-Validation of Total Sample Results ...340

Summary of Factor Analytic Results 347

C. Digit Span Analysis 351

V. Discussion 359

A. General Recommendations 359

B. Age Trends in Number and Definition of Factors 365

C. Individual Subtest Anomolies 370

Similarities and Picture Completion 371

Digit Span 379

Picture Arrangement and Mazes 385

D. Limitations of the Study 386

E. Summary Statement 391

Reference Notes 394

References 395

List of Tables

Table Page

1. Final Sample Distribution by Village, Sex, and Age ..248

2. Split-half reliability coefficients:
N.W.T. and U.S. Norming Samples 260

3. LISREL Matrices in the Factor Analytic Model 273

4. Clinical Models to be Tested:
Wechsler Scales as Two Factors 283

5. Clinical Models to be Tested:
Kaufman Three-factor Model 284

6. Clinical Models to be Tested:
Bannatyne Four-factor Model 286

7. Subtests with Distributions Deviating from N(10,3) ..288

8. Equality of Covariance and Correlation Matrices 292

9. Subtest Intercorrelations for Age Pool 7-8 Years.....293

10. Subtest Intercorrelations for Age Pool 9-10 Years ...293

11. Subtest Intercorrelations for Age Pool 11-12 Years ..294

12. Subtest Intercorrelations for Age Pool 13-14 Years ..294

13. Subtest Intercorrelations for Total Sample 295

14. Summary of ML Analyses: Total Sample 296

15. Principal Factor Analysis: Total Sample
Promax Solution for Two Factors 298

16. Maximum Likelihood: Total Sample
Final Model for Two Factors 299

17. Principal Factor Analysis: Total Sample
Promax Solution for Three Factors 301

18. Maximum Likelihood: Total Sample
Final Model for Three Factors 304

19. Principal Factor Analysis: Total Sample
Promax Solution for Four Factors 306

20. Maximum Likelihood: Total Sample
Final Model for Four Factors 308

21. Summary of ML Analyses: Age Pool 7-8 Years 310

22. Principal Factor Analysis: Age Pool 7-8 Years
Promax Solution for Two Factors 312

23. Maximum Likelihood: Age Pool 7-8 Years
Final Model for Two Factors 313

24. Principal Factor Analysis: Age Pool 7-8 Years
Promax Solution for Three Factors 314

25. Maximum Likelihood: Age Pool 7-8 Years
Final Model for Three Factors 315

26. Summary of ML Analyses: Age Pool 9-10 Years 317

27. Principal Factor Analysis: Age Pool 9-10 Years
Promax Solution for Two Factors 318

28. Maximum Likelihood: Age Pool 9-10 Years
Final Model for Two Factors 320

29. Principal Factor Analysis: Age Pool 9-10 Years
Promax Solution for Three Factors 321

30. Maximum Likelihood: Age Pool 9-10 Years
Final Model for Three Factors 323

31. Summary of ML Analyses: Age Pool 11-12 Years 325

32. Principal Factor Analysis: Age Pool 11-12 Years
Promax Solution for Two Factors 327

33. Maximum Likelihood: Age Pool 11-12 Years
Final Model for Two Factors 328

34. Principal Factor Analysis: Age Pool 11-12 Years
Promax Solution for Three Factors 329

35. Maximum Likelihood: Age Pool 11-12 Years
Final Model for Three Factors 330

36. Summary of ML Analyses: Age Pool 13-14 Years 333

37. Principal Factor Analysis: Age Pool 13-14 Years
Promax Solution for Two Factors 334

38. Maximum Likelihood: Age Pool 13-14 Years
Final Model for Two Factors 336

39. Principal Factor Analysis: Age Pool 13-14 Years
Promax Solution for Three Factors 337

40. Maximum Likelihood: Age Pool 13-14 Years
Final Model for Three Factors 338

41. Principal Factor Analysis: Age Pool 13-14 Years
Promax Solution for Four Factors 339

42. Cross-validation of ML Result for Total Sample
Age Pool 7-8 Years..................................341

43. Cross-validation of ML Result for Total Sample
Age Pool 9-10 Years.................................343

44. Cross-validation of ML Result for Total Sample
Age Pool 11-12 Years................................344

45. Cross-validation of ML Result for Total Sample
Age Pool 13-14 Years................................345

46. Number of Factors Suggested by PC and ML Methods 348

47. Subtest Correlations With Digit Span Forward
and Backward: By Age Group 353

I. Introduction

The present research involves an examination of the validity of the Wechsler Intelligence Scale for Children (WISC-R) for diagnosis of the educational difficulties experienced by children in Canada's eastern arctic regions. The present chapter describes the context in which the need for this research arose, the author's assumptions regarding the ethical demands associated with such research, and establishes the limits of the issues specifically raised in the report. This chapter also serves as an advance organizer for a review of the literature pertaining to topics in the areas of assessment, intelligence theory, test bias, and factor analysis.

A. Setting for the Research

Watters (1980, in press) described a set of challenges facing educators in the Northwest Territories (NWT) and proposed a series of assessment and remediation services to address those challenges. The major problems listed by Watters were age-grade retardation, hearing impairment, inadequate school readiness skills, dropping out, sporadic attendance, language conflict, and cultural conflict. Some of these challenges are described in more detail in Chapter II.C. The assessment proposals offered by Watters included evaluation of the cognitive abilities of children in the NWT school population, assessment of the local needs for educational resources, and analysis of the special education services currently in operation.

Watters (1980) perceived a need to distinguish children requiring special education from those requiring remedial education. Watters defined special education as the manipulation of the educational environment to counteract conditions which interfered with the intellectual, emotional, physical, or communicative functioning of the child. Mental retardation, learning disabilities, neurological and emotional disorders were listed as examples of such handicapping conditions. Remedial education was defined as the provision of extra assistance to compensate for educational disadvantages which were not due to a specific physical or learning disability. Low achievement due to sporadic school attendance, hearing loss, adolescent alienation, and inadequate mastery of the first language were listed as examples of problems requiring remedial education.

Watters (in press) stated that the cognitive tests which had been used most effectively by NWT psychologists were the WISC-R, the Bender Motor-Visual Gestalt Test (commonly referred to as the Bender), and the Goodenough Draw-A-Man (DAM) and Draw-A-Woman (DAW). The absence of local norms for these tests was perceived as an impediment to the development and provision of diagnostic services. A research team led by Mulcahy administered the test to samples of school children aged 7 years., 0 months to 14 years., 11 months in the Keewatin and Kitikmeot regions of NWT. Age norms were provided for each test from those scores (Mulcahy & Watters, 1982). The present study uses the scaled scores for that norming sample to test the construct validity of the WISC-R for Inuit children.

B. Value and Assumptions of the Present Study

Among its standards for the development and publication of educational and psychological tests, the American Psychological Association (1974) stated that it is essential that a test author indicate the evidence substantiating any proposed theoretical or construct interpretation of the test. That association (APA) defined a construct as "a dimension understood or inferred from (the test's) network of interrelationships" (p. 29) and construct validity as the degree of confidence which may be placed in interpretation of the test as a measure of that dimension. Scores on the WISC-R have been interpreted according to various construct definitions, based on research with samples which were drawn predominately from schools in the U.S.A. Generalization of the evidence for these constructs to NWT children cannot be assumed, for reasons which are discussed in depth throughout Chapter II. The present study examines the validity of three sets of construct definitions, or interpretive models, which have been proposed for the 12 subtests of the WISC-R.

Evaluation of an instrument such as the WISC-R is influenced by the assumptions and priorities of the researcher. Phillips (1980) has warned the consumers of evaluation reports to "pay particular attention to the linking value premise or premises that must exist" (p. 24). In the interest of consumer awareness, some of the author's assumptions and premises are listed below.

1. The responsibility for determining the validity of an interpretive model of the test lies with the agency or tester advocating that interpretation. In other words, a psychologist would be negligent in assuming that a given model is valid for NWT children until proven otherwise.
2. Each interpretation inferred from a child's WISC-R profile must be empirically supported for the NWT population, whether the inference relates to underlying cognitive ability or to future performance. Therefore, the present study is only an initial step in the process of validating the WISC-R.

The limitations of the present study are discussed in Chapter V, since understanding of many of the theoretical and methodological constraints requires awareness of the issues discussed in Chapter II. A brief outline of those issues is provided below.

C. Advance Organizer to the Literature Review

The overall aim of the literature review is to inform the reader of a set of issues pertaining to the validity and generalizability of inferences regarding cognitive processes and/or capacity from scores on norm-referenced tests, particularly the WISC-R, and the role which factor analysis may play in assessing construct validity.

Chapter II.A begins by establishing that current uses of normative intelligence tests in psychological or educational assessment involve inferences of underlying cognitive ability traits, which influence scores on sets of individual tests. Intelligence itself is such a trait, or hypothetical construct. The scientific usefulness of constructs is discussed, leading to the introduction of factor analysis as a means of operationalizing a set of constructs. Several dominant theories of intelligence have been derived and/or supported from factor analytic research. A brief description of some of these theories and the test batteries they spawned reveal that a given set of data may support more than one arrangement of factors when subjected to different extraction and rotation methods. Factor analysis may organize tests into clusters which account for the maximum amount of common variance, but does not indicate the nature of the trait underlying that common variance. Consequently, the role of factor analysis in validating a construct interpretation of a test is to determine whether the interpretation is tenable, i.e., to determine whether the test clusters with tests believed to measure the same construct. Whether the factor is a measure of the construct of interest must be assessed by other methods.

Chapter II.B reviews the literature on the validity of construct interpretations of the WISC-R. Substantial, if not unequivocal, evidence is cited for Wechsler's organization of the subtests into the Verbal and Performance Scales and for Kaufman's three-factor interpretive model. Three of Bannatyne's four categories received empirical support in the studies reviewed, although his Acquired Knowledge factor was not supported. Although these factors have been replicated by several methods and with several samples, there are disagreements regarding the nature of the constructs represented. A review of some of the experimental literature with tests such as Digit Span revealed that abilities which are assumed to be necessary for success on the task (and perhaps correctly so) may not be a factor in the variance of test scores for a particular sample. Extraneous variables, such as English fluency and familiarity with English phonemes, may be a factor in the variance of test scores for other samples. Analysis of profiles of individual subtests was not supported by the studies reviewed in this report, so that the clinical utility of the WISC-R appears to be dependant on the validity of the various factor models. Although the results of studies of the criterion-related validity of these proposed factors have been promising, the constructs which underlie the factors have not been clearly identified.

Chapter II.C reviews the literature pertaining to cognitive assessment of Native North American children. Trends on the WISC-R and other tests are described and the types of inferences made from these trends are critiqued. Overinterpretation of test and factor scores has been the rule in most of the studies reviewed, as reflected in inferences about the cognitive processing strategies and capacities of Native children in general. Research is reviewed which demonstrates the effect on test scores of such extraneous factors as ear infection or familiarity with the English sound system. Given the array of factors which may effect the test scores of Inuit children, interpretation of the factors derived from the present research must be hypothetical.

In Chapter II.D, the literature review switches to a technical discussion of the reliability of factors, particularly where these factors have been rotated to maximize their agreement with an hypothesized factor model. The power of factor analysis to test the construct definition of a test or test battery is examined in view of the technical concerns. Maximum likelihood factor analysis (MLFA) was chosen as the method which allowed the most rigorous and direct test of the factor models. The strengths and weaknesses of MLFA are discussed in detail.

MLFA involves a series of decisions regarding the setting of parameters in the factor model, which are sequentially adjusted until all hypothesized models have been confirmed or rejected. The method and results of analysis are presented in extensive detail to inform the reader of the steps leading to the acceptance of a factor model for a given age level. Summaries of the results for each age level are provided following discussion of those results. The discussion in Chapter V focuses on the trends for support of the factor models across the age groups. Where a model is rejected, tentative explanations are discussed.

II. Review of Literature

A. The Importance of Construct Validity to Assessment

The present section of the literature review describes the relationships among several concepts which will be explored in more depth in subsequent sections. Intelligence is defined as a hypothetical construct and the assessment of intelligence as a procedure to generate inferences about a child's behavior in relation to that construct. Factor analysis is presented as one method of testing the viability of such constructs for interpretation of a set of test scores. The relationship of theories of intelligence to the factor analytic methods employed by the theorists is explored, demonstrating the limitations of factor analysis for explaining the constructs derived from the method's application. Test bias is introduced as an issue pertaining to the generalizability of inferences from test scores across populations. This section is concluded with the assertion that the validity of present applications of intelligence testing is limited by the evidence for the viability of the construct of intelligence implied by a given test's interpretation, and by the evidence for that construct as a factor underlying the test in question.

Assessment of Intelligence: Purposes and Assumptions

The administration of educational and psychological tests is applied to several problems in the management of instruction. Horn (1979) categorized these test applications as sorting, grading, and monitoring. Intelligence tests are usually applied only to the task of sorting children according to such criteria as aptitude or school readiness. Criterion-referenced tests are usually applied to grading, i.e., evaluation of children's present achievement, and to monitoring of the effectiveness of educational programs. Salvia and Ysseldyke (1981) listed screening, classification, planning of educational programs, program evaluation, and assessment of individual progress as purposes of psychoeducational assessment. This list corresponds closely to Horn's, although Salvia and Ysseldyke noted that intelligence tests are often used for the planning of individual educational programs. The latter authors distinguish between the sorting tasks of screening to identify children with special educational needs and classification of such children into diagnostic categories. The assumptions underlying the valid use of tests for any of the purposes are that: the tester is skilled in the administration of the test(s) used; the child's score will reflect some random error; the child has been exposed to the same acculturating environmental influences as the children represented by the test norms; that the ability or behavior to be assessed has been adequately assessed by the test items; and that present behavior is observed, whereas future behavior may only be inferred (Salvia & Ysseldyke, 1981).

The authors listed above have described the various assessment procedures as responses to demands for information, all of which might be required in making decisions about a child's education. Swanson and Watson (1982) have categorized four models of assessment, which they claim underlie the type of information sought and the means used to acquire it. The Construct, or Attribute, Model assumes that children may be characterized by their placement on the continuum of an ability dimension, such as verbal reasoning, sequential memory, or nonverbal intelligence. Academic difficulties are viewed as effects of deficiencies in one or more such abilities. Diagnostic prescriptive teaching, which consists of assessment to determine the cause of a child's learning handicap(s), identification of preferred learning styles and strategies, and remediation according to that pattern of strengths and weaknesses, is conceptually linked to the the Construct Model of assessment (Swanson & Watson, 1982). This model is reflected in the procedures for determining the presence of a specific learning disability, as established within U.S. federal law (Salvia & Ysseldyke, 1981). A child may be diagnosed as having a specific learning disability if there is a severe discrepancy between intellectual ability and achievement in one of: oral expression; listening comprehension; written expression; basic reading skill; reading comprehension; mathematics calculation; or mathematics reasoning. Implicit in these diagnostic criteria is the concept of one or more underlying intellectual abilities, whose expression may be inhibited by poor task strategies, perceptual difficulties, or some other cognitive weakness.

The Construct Model was critiqued by Swanson and Watson (1982). These authors argued that the expression of abilities in the testing situation may not generalize to other situations. However, the relationship of test scores to performance on some criterion measure appears to be an empirical question which may be examined for each test of interest. Other criticisms of the assessment model have included charges that the theoretical rationale for interpretation of test score profiles is not well articulated (Drenth, 1972; Swanson & Watson, 1982) and that the tests may measure different abilities for different groups (Drenth, 1972). Swanson and Watson also questioned the assumption that profiles of strengths and weaknesses can be translated into effective teaching strategies. This skepticism is supported in reviews of the literature on the effectiveness of differential diagnosis and prescriptive teaching from ability profiles (Arter & Jenkins, 1979; Ysseldyke & Mirkin, 1982). However, that model of assessing and remediating educational difficulties was favored by a vast majority of the special educators surveyed by Arter & Jenkins (1979).

The other models described by Swanson and Watson (1982) are the Functional, Ecological, and Decision-making Models.

Assessment in the Functional Model involves task analysis of the behaviors and skills required of the child and identification of the environmental antecedents of those behaviors. Testing is criterion-referenced, rather than norm-referenced. The Ecological Model focuses on the social systems within which the child operates, such as the school and the family. The Decision-making Model is more similar to the Construct Model, in that the assessment is focused on the traits which underlie the child's performance. Swanson and Watson distinguished the Construct and Decision-making Models by describing them as passive and active models, respectively. Assessment in the latter model is described as a sequential process of testing, forming hypotheses based on the results, and testing these hypotheses with further testing, observations, case conferences, etc. Tucker's (1977) proposed assessment strategy is an example of a Decision-making approach. Assessment begins with analysis of school records and observational data, followed by assessment of language dominance, sensory, motor and linguistic abilities. The focus of assessment then switches to adaptive behavior and medical and developmental health. Assessment of personality traits and intelligence are the final stages in the procedure. At the conclusion of each assesssment stage, the parents and school personnel meet to decide whether to conduct further assessment, retain the child in the regular classroom, or prescribe some form of remedial education.

Swanson and Watson appear to favor the Decision-making Model over the other models they described. Drenth (1972) has noted that the role of the test is more clearly identified in active decision-oriented assessment procedures. However, Drenth also noted that tests may be used incorrectly for any given decision. When hypothetical constructs are assumed to underlie test performance, inferences about a child's placement on the dimension of that construct and subsequent decisions about the child's education are limited by the validity of those constructs and the test's ability to measure them.

Messick (1980) conceived of test validity as an overall evaluative judgement of the inferences drawn from test scores. This judgement is based on: an inductive summary of the discriminant and convergent validity research supporting interpretation of the test as a measure of a particular construct; an appraisal of the values implied by that test interpretation; an evaluation of the rationale for the relevance and utility of scores on that construct to the testing application; and an appraisal of the potential social consequences of that application of the test. The use of an overall validity measure is strongly opposed by some authors and others question the validity of the concept of construct validity. The following section examines that concept, with discussion of its theoretical and operational definitions.

Definition and Measurement of Construct Validity

Construct validity has been defined as the degree to which a test can be "interpreted as a measure of some attribute or quality which is not 'operationally defined'" (Cronbach & Meehl, 1955, p. 282), i.e., some attribute of which no one criterion is an adequate measure. The attribute is defined by a theory regarding its relationships with other attributes, some of which must be directly observable. Definition of the construct is clarified by expanding knowledge about the network of relationships among these attributes, i.e., the nomological network. The nomological network can be investigated through the analysis of correlation patterns, differences in group means, test item analysis, factor analysis or any other procedure for generating or testing hypotheses about the network's structure. Where relationships hypothesized by the theory are demonstrated, and where there is an absence of findings which threaten the hypothesized structure of the network, there is evidence for construct validity. Where support for the theoretical definition of the construct is not available, several interpretations of the findings are possible: the test may not measure the construct hypothesized; the theory about the construct's relationship with other observed variables may have been incorrect; or the experimental design may have been inappropriate to test the hypothesis. Cronbach and Meehl argued that the burden of proof is on the test interpreter to demonstrate that the fault lay with the theory in such cases, rather than with the test.

The concept of construct validity has been sharply criticized by several researchers. Bechtoldt (1959) stated that a variable which is, by definition, not operationally defined, is not scientifically useful. He further argued that Cronbach and Meehl (1955) had confused the meaning and significance of a construct. Bechtoldt defined a construct's meaning as its operational definition and its significance as the strength of its relationships with other variables, i.e., its adherence to a set of theoretically defined laws. Cronbach and Meehl had inferred meaning from significance.

In his discussion of the construct validity of cognitive tests for use with minority groups, Jensen (1980) has offered an argument similar to Bechtoldt's (1959). Jensen stated that equivalent construct validity across groups could not be proven or disproven. Only the equivalence of a test's relationships with observable criteria could be demonstrated. Cress (1974) has warned that cognitive tests should not be interpreted as indicators of cognitive capacity, but are useful as predictors of future academic success within the public school system in the United States. Jensen's discussion appears to reflect a reaction to the vagueness of the concept and measurement of construct validity, whereas Cress appears to be reacting to the consequences of overinterpretation of test results for the educational futures of the children tested. These concerns are justified by much of the current research and application of cognitive testing, as reviewed throughout this chapter. However, the restriction of validity research to the test's power to predict individual criteria, or sets of individual criteria, is not without conceptual and methodological difficulties.

Goldman and Slaughter (1976) provide an example of misleading validity indices resulting from poor criterion selection. The Scholastic Aptitude Test (SAT) had a low correlation with college Grade Point Average (GPA) when applied to the scores of students across a wide range of college disciplines. However, the correlations calculated for SAT with grades in individual courses were much higher. Goldman and Slaughter were able to demonstrate that differences in the grading patterns across disciplines, plus differences in the academic ability of students entering the various disciplines, lowered the correlation of SAT scores with the composite GPA.

The value of a criterion for predictive validity is also affected by the events occurring between the administration of the predictor test and the criterion measure. Cleary, Humphreys, Kendrick, and Wesman (1975) argued that the use of standardized tests to predict school performance implied the assumption that there would be few radical changes in school curriculum or instructional methods. A test which is valid for predicting children's performance in mainstream school classrooms may not be a valid predictor of performance in more innovative or experimental school settings (Flaugher, 1978).

In his discussion of the prediction of occupational success, James (1973) noted that current models for measuring criteria for predictive validity do not offer much guidance for the selection of criteria measures. He criticized the use of a single criterion measure, challenging the implicit assumption that a single factor underlies occupational success. James argued for the measurement of several criteria for test validity, including measures of both personal and organizational effectiveness.

Some of the early discussions on construct validity may be faulted for their lack of explicit operational criteria for determining construct validity. However, advocacy of a bivariate approach to test validity, concerned strictly with accuracy of prediction of academic achievement under a particular educational treatment, begs the questions of which criteria are important to predict and under which educational treatments. Given that current cognitive assessment procedures which employ norm-referenced tests involve assumptions about cognitive abilities which underlie school achievement and responses to academic treatments, it seems important to assess the evidence for the validity of those assumptions. The criterion-validity problems which may arise in construct validity research would be controlled most effectively by simultaneously and systematically examining as many criteria as possible, i.e., looking at as large a slice of the nomological network as can be managed with available samples. Several methods for investigating construct validity have been proposed, including factor analysis and integration of past research on the criterion-related validity of the test (Cronbach & Meehl, 1955; Messick, 1980). Royce (1963) has stated that factors (particularly higher-order factors) are operational definitions of hypothetical constructs. Factor analysis is not the only multivariate technique available to generate or test hypotheses about the pattern of a test's relationships with other variables. However, it has been applied to the definition of constructs of a wide variety of cognitive abilities, particularly the construct of intelligence.

The following section examines theories of intelligence which are based on the factor analytic model. Some concepts pertaining to that model are introduced in the following discussion, but an elementary understanding of factor analysis is assumed. Readers who are unfamiliar with factor analysis are referred to Mulaik (1972) or a more concise introduction by Cooper (1983).

Factor Theories of Intelligence

The description of a selection of factor theories of intelligence is included to inform the reader of some of the differing concepts of the nature of intelligence factors. It is also intended to to illustrate the close association between the theorists' concepts of intelligence and the particular factor analytic methods they employed. The list provided here is certainly not exhaustive of factor theories of intelligence, but is sufficient to highlight some of the major issues in that area of research.

The first factor analytic model of intellectual abilities is attributed to Charles Spearman. Spearman's two-factor theory (1904) states that measures of such abilities were influenced by two factors: *g*, which represented a general ability required for all tasks; and *s*, a specific factor representing skills unique to that test. The specific factors are uncorrelated with each other. Spearman (1927) later refined his theory to elaborate on the definition of *g*, which he described as the ability to perform the eduction of relations and correlates, i.e., to understand associations among objects, events, or ideas and form hypotheses concerning those relationships. Spearman's theory failed to account for the fact that intercorrelations within groups of similar tests were often larger than their individual correlations with *g* would predict (Brody & Brody, 1976). Observation of such clusters of tests led to the theoretical formulation of group factors.

Thurstone (1938) identified seven group factors, which he labelled primary mental abilities, from the scores of college students on a large battery of tests. These group factors were labelled Spatial, Perceptual, Numerical, Verbal Relations, Words, Memory, and Induction. The evidence for two additional factors, named Reasoning and Deduction, was considered tentative. Thurstone constructed the Primary Mental Abilities Test (PMAT) with tests from his battery, from which scores on the first seven factors were calculated and interpreted.

Thurstone rejected the concept of a general factor. Each test score was believed to be influenced by ability on one group factor plus ability on the skills unique to the test. Adherence to this condition by a battery of tests was labelled simple structure. Operationally, simple structure is defined as the presence of substantial factor loadings on only one factor per test. Thurstone (1938) achieved this organization of loadings by graphical rotation of the axes in plots of tests' loadings on pairs of factors.

The factoring methods chosen by Spearman and Thurstone reflected the respective theories of those authors. Spearman's law of tetrad differences was only applicable to a general-factor theory. Thurstone's use of centroid factoring procedures facillitated the extraction of multiple factors. The principle of simple structure is not consistent with the presence of a general factor in the factor matrix, since that factor would have no nonsalient loadings, resulting in the presence of complex variables with tests loading on both the general factor and a group factor (Brody & Brody, 1976). Brody and Brody have suggested that hierarchical theories of mental abilities allow a synthesis of the ideas of Spearman and Thurstone.

More recent theories of intelligence have hypothesized a hierarchical organization of mental abilities, with factors at each level exerting an influence upon test scores. This set of influences may be expressed mathematically as

$$Z_j = a_jG + b_{j1}H_1 + b_{j2}H_2 + c_{j1}F_1 + . . . + U_j \quad (II.1)$$

where a score Z on test j is a function of the person's ability on the general factor, G, a set of K higher-order group factors, H_k, a set of M primary group factors, F_m, and the ability (or abilities) unique to test j (Mulaik, 1972). The coefficients a_j, b_{jk}, and c_{jm} are loadings for the test on the various factors at the respective levels. Hierarchical factor analysis is generally executed by factoring the correlation or covariance matrix for the observed test scores, followed by oblique rotation of the resulting primary factors. The correlation matrix for the primary factors is subsequently factored and rotated. The number of levels in the hierarchary will be a function of the theory underlying the research, the number of test scores observed for each subject, the size of correlations among factors at lower levels, and other results which may suggest the influence of higher order factors. Simple structure may be retained by the matrix of loadings on the oblique primary factors, while the effect of more general abilities are reflected by the loadings on the higher-order factors. The nature of mental abilities at the various levels are still a matter of debate among theorists who espouse hierarchical models. Two such models are briefly described below.

Vernon (1965) proposed a theory in which all mental abilities were affected by a general intellectual ability, labelled *g*. Subsumed under this factor were the major group (second-order) factors: *v:ed*, which underlied skills requiring verbal reception and expression; *k:m*, which involved psychomotor and perceptual abilities; and a factor which Vernon tentatively labelled as *I*, for inductive reasoning ability. In turn, *v:ed* abilities included the minor group (primary) factor of creative abilities and specific factors such as reading, linguistic and clerical skills. Primary factors reflecting spatial and mechanical abilities were subsumed under the *k:m* factor. Vernon's theoretical structure does not assume simple structure for the loadings of tests on primary factors. Mathematical abilities were portrayed as a function of a combination of all three second-order factors, while skills such as reading were thought to reflect both *v:ed* and perceptual speed, a primary factor in the *k:m* group.

Cattell has proposed a hierarchical theory of intelligence in which two general factors operate upon the primary factor and test scores (Cattell, 1971; Horn, 1968). Crystallized intelligence (*Gc*) is a structure of simple and complex concepts and skills, acquired through acculturation. Vocabulary, social skills, and verbal comprehension are examples of *Gc* factors. Fluid intelligence (*Gf*) is a set of intellectual abilities which are predominantly gained through incidental learning, rather than acculturation, and considered to be more dependent upon neurological integrity. Memory, inductive reasoning, and comprehension of figural relations are examples of *Gf* abilities. Horn (1968) noted that the predominant influences of physiological and acculturative factors upon *Gf* and *Gc*, respectively, are a matter of degree, as both of the general factors are influenced by acculturation and physiological variables. Scores on measures of Formal Reasoning (a primary factor in their intelligence model) appeared to be affected by both of the general factors.

The major differences between the hierarchical theories and results of Vernon and those of Cattell and Horn appears to be the method of factor rotation and the number of general factors identified. Vernon's *g* is conceived as similar to Spearman's *g*, whereas the two general factors of Cattell and Horn reflect the idea that types of intelligent behavior may be differentiated by the modes in which they were learned and the complexity of neurological networks supporting them. The distinction between abstract and practical abilities inherent in Vernon's definition of *v:ed* and *k:m* are not applicable to Cattell and Horn's theory (Horn, 1968). The two theories also differ in their interpretation of the higher order factors underlying such primaries as Mechanical Abilities and Visualization. Horn suggested that the differences between these theories reflect Cattel and Horn's use of objective modes of factor rotation, with simple structure as a criterion, in contrast to Vernon's use of more subjective rotation techniques and criteria. These two theories are neither the only dominant hierarchical theories nor the first. Comparison between these theories serves to make the point that the use of hierarchical factor analysis has helped to synthesize the theories of Spearman and Thurstone without resolving the debate regarding the importance or nature of higher-order mental abilities.

Jensen has proposed a hierachical theory of intelligence which differs from those of Vernon or Cattell in that the general factor is portrayed as an ability trait which is more developmentally advanced than the lower-order factor. These factors are labelled level II and level I intelligence, respectively (Jensen, 1976, 1980). Level I intelligence is comprised of rote learning and memory ability, while level II intelligence reflects an ability for abstract thought and performance of tasks which require mental transformations of concepts or object representations. For example, Jensen (1980) suggested that simple arithmetic calculation is a level I ability, whereas arithmetic problem-solving is a level II ability. A person may be low functioning in regard to level II intelligence, while performing in the average range on level I. Jensen (1980) uses the terms *level II intelligence* and *g* interchangeably, and has explicitly identified the concept of intelligence with *g*.

Jensen's operational definition of *g* is the unrotated first principal component for a heterogenous collection of cognitive tests (1976, 1979, 1980). His rationale is that this component accounts for the greatest possible amount of all the test variance. Tests which have high loadings on the component are considered to be good measures of *g*. He has cited Raven's Progressive Matrices, verbal analogies, and series completion tasks as examples of tests with high *g* loadings (Jensen, 1980).

Jensen has admitted that loadings on the first principal component will be dependent upon the collection of tests included in the analysis, but argued that restriction of the tests to those in the domain of mental abilities would result in a definition of *g* which resembled level II intelligence. However, his method of identifying *g* assumes the presence of a general factor, without testing that assumption. Carroll (1981a, 1981b) reanalyzed a set of reaction-time data which Jensen (1979, 1980) had reported, finding three orthogonal factors through principal factor and maximum likelihood analysis. The orthogonality of the factors precluded the use of hierarchical analysis. Carroll (1981b) also reanalyzed Primary Mental Abilities Test data by hierarchical analysis. Carroll found a general factor but less than half of the common variance was in the higher--order domain. This finding contradicted Jensen's (1980) claim that the general factor contributes more to the variance of the PMA subtests than is contributed by the primary factors. Consequently, it challenges Jensen's assertion that the component which accounts for the greatest amount of variance may be identified as *g*, or level II intelligence.

Jensen's attempts to validate his *g* construct have assumed adherence to his theory that the low average IQ scores obtained by U.S. black children are a function of racial differences in capacity of *g* (Jensen, 1976, 1980). His argument is that, since blacks have less *g* ability, on the average, than whites and since cognitive tests with high loadings on the first principal component tend to distinguish these groups, high loadings on the first component reflect *g*. Thus, Jensen's interpretation of tests as measures of level I or level II intelligence, plus his suggestion that education for black children focus on level I skills (Jensen, 1971) are based on circular logic in which his cognitive theory and his methods are used to validate each other.

While other researchers have debated the value and validity of group and general factors, Guilford has devised a theory which views intelligence factors as cross--classifications of specific abilities (Guilford, 1967; Guilford & Hoepfner, 1971). Abilities are classified according to the mental operation required, the content areas represented, and the product, i.e., the type of information gained by exercising the ability. Operations include cognition, memory, convergent and divergent production, and evaluation. Contents include symbolic, semantic, figural, and behavioral task materials. Products include information about units, classes, transformations, systems, relations, and implications. Guilford and Hoepfner stress the point that the model is not hierarchical, i.e., classifications are not subsumed under each other. This model proposes that 120 mental abilities may theoretically be identified (the number of possible combinations from five operations, four content areas, and six products). Tests used in the Structure-of-Intellect (SI) research program are each designed or selected to measure only one ability. For example, a verbal-analogies item could tap the ability to perform cognition of semantic relations, whereas an analogies item with geometric content would measure ability in cognition of figural relations.

Guilford and Hoepfner (1971) conducted factor analytic research in support of the SI model. The number of factors extracted from the test battery for each study were decided by the number expected according to the theory. Factor matrices were rotated to maximum congruence with the factor structure proposed by the SI model. Objective measures or criteria for the number of factors or their rotation were evaluated on their adherence to the model, rather than the reverse procedure. Guilford and Hoepfner noted that rotation to simple structure resulted in confirmation of only 32% of the SI factors hypothesized across their series of studies.

In contrast, rotation toward congruence with their model resulted in 92% agreement. Their logic of analysis, which reflects overconfidence in their ability to design tests which measure exactly what they wish, does not test their model. Their conclusions are further threatened by problems inherent in the use of subjective methods of rotation, which are discussed in Chapter II.D of the present report. However, the SI designations for several standardized tests of intelligence are provided as interpretive aids in some major texts on psychological testing (e.g., Kaufman, 1979a; Sattler, 1982).

Construct Definition Across Age Spans

The authors of some of the factor models discussed above have noted that the organization of intellectual abilities may vary across age groups, and that factor models should be validated at every age to which the models are to be generalized (Horn, 1968; Thurstone, 1938). Garret's (1946) differentiation hypothesis states that abstract intelligence develops from "a fairly unified and general ability to a loosely organized group of abilities or factors" (p. 373). His discussion implies that abstract intelligence refers to an ability to manipulate symbols and deal with abstract concepts. His conclusions were based on cross-sectional research and trends detected in a review of published factor analytic studies with various age groups. Garrett suggested that the general factor, which he believed was reflective of linguistic ability, declined in importance as children reached adolescence and throughout early adulthood. Moderate support was provided for this hypothesis through comparison of the factor structures of subjects aged 10 to 12 years, 18 to 20 years, and 45 to 60 years using a battery of 14 standardized tests (Leinert & Croft, 1964). Leinert and Croft also provided moderate support for the dedifferentiation hypothesis, which states that the influence of general intellectual ability increases in late adulthood. Guilford (1967) has noted a number of methodological problems in the research pertaining to changes in the factor structure of intelligence over the age span. He argued that it is difficult to find tests or test batteries which are suitable across the age span (e.g., which have an adequate range of item difficulty levels to obtain similar variances for the different age groups). Guilford also noted that many of the studies published at that time had compared samples at age 14 years and above, where most of the primary factors evident in adulthood had already been identified. This practice may have masked a differentiation effect which would have been evident with younger age groups. Chapter II.D of this report explores some technical issues pertaining to the comparison of factors across groups.

Limits to Factor Interpretation

The theories of intelligence reviewed above were closely linked to the factor analytic methods used to explore or test them. Intelligence has been defined as a general trait influencing ability level on all cognitive tasks, a structure of several levels of cognitive ability, or sets of isolated cognitive abilities. Sternberg (1980, 1981) has faulted factor analysis for being unable to disconfirm too many theories of intelligence. He argues that factor analysis clusters products rather than cognitive processes, noting that Guilford's SI model is the only factor model of intelligence which distinguishes between task content and cognitive operations. The claim that factor analysis does not provide information on cognitive processes has been supported by Brody and Brody (1976) and Carroll (1981a). Carroll suggested that, although the absence of a correlation between two tests probably suggests differences in processing, the presence of a correlation may be due to similarity of task content, cognitive processing, or other variables. Messick (1972) argued that factor analysis had identified several important cognitive traits, or latent constructs, but was unable to provide functional linkages between the constructs. In other words, scores on mental ability factors do not provide information on the manner in which a child expresses those abilities in a complex task.

The above criticisms of factor analysis did not lead their authors to argue for the abandonment of that method in construct validation research. Sternberg (1980) suggested that factor scores are useful for the prediction of future performance on similar tasks. When the goal of test interpretation is diagnosis of educational disabilities or handicaps, Sternberg suggests that factor analysis should be followed by an analysis of the components, or information--processing operations involved in responding to a test item. Componential analysis generally consists of measuring correlations between psychometric test scores and skill in the behaviors believed to be components of performance on the test. The component variables are often reaction-time scores on simple tasks, such as visual discrimination tasks. Sternberg (1981) has since noted that many of the components described in his own and similar research are also hypothetical.

Embretson (1983) has attempted to clarify the concept of construct validity and its measurement by identifying two types of construct validity, with associated methods for measurement. Construct representation refers to the "theoretical mechanisms that underlie task performance" (p. 180). These mechanisms, or constructs, correspond to Sternberg's (1980, 1981) components. Embretson noted that ability level on a component which is essential for a correct response on a test item will not necessarily be associated with variance on the test score. The population of interest may not vary systematically on that component ability. Embretson proposed a method for assessing construct representation, labelled Multicomponent Latent Trait Modeling (MLTM), which integrates aspects of Sternberg's componential analysis with latent trait methods of item analysis. The pattern of the test's relationship to other measures constitutes nomothetic span, the second type of construct validity. The concept of nomothetic span corresponds with Cronbach and Meehl's (1955) nomological network. Factor analysis is appropriate for addressing questions regarding nomothetic span, but inappropriate for assessing a test's construct representation.

Messick (1972) argued that a complex task involved a sequence of subtasks, much like Sternberg's (1980, 1981) components. Ability on a given component might correlate with ability on the overall task until the component was overlearned by most subjects in the sample or until they had developed a strategy to overcome deficits on the component. Therefore, the constructs underlying test scores could differ across groups with varying amounts of practice or exposure to the task. As an example, Messick cited Fleishman's (1960, 1967; Fleishman & Hempel, 1954) demonstration that, although visualization ability was related to early performance on a set of psychomotor tasks, the strength of that relationship decreased with practice in the latter tasks. Motor speed and coordination became better predictors of psychomotor performance as practice time increased. Messick (1972) called for the merging of factor analytic and experimental methodologies for construct validation. He suggested a multivariate approach, where independent variables, such as practice, test instructions, or materials would be manipulated to determine their effect on factor patterns and factor scores. Royce (1963) had earlier suggested that factor scores could be effectively used as dependent or independent variables in learning research.

The argument that a test may measure a different construct at different stages of learning implies that the test may measure different constructs for groups of people from dissimilar learning environments (Messick, 1972). Horn (1968) had recognized this problem and its implications for operationally defining crystallized and fluid intelligence. If the acculturative practices of two populations differ, a test's relationship to *Gc*, which reflects acculturation, and to *Gf*, will differ across the two populations. Therefore, inferences about the construct underlying a test must be validated for every new population to which the test is applied. The invalid generalization of inferences about test scores to members of a new population comprises biased use of the test. A more complete definition of test bias is provided below, with a discussion of its implications for assessment practices.

Test Bias: Definitions and Significance

The preceding discussion on inferences regarding hypothetical constructs identified test validity as a function of the degree to which test scores actually reflect the effect of that construct. Other types of inferences are made from scores on psychological or educational tests. Predictions about future performance on some other measure, such as academic achievement, may be derived from present scores on an aptitude test without naming a construct as a factor underlying either performance. Thus, the validity of a number of intended inferences may be investigated and some may be rejected while others are supported. As there are many types of test validity, Flaugher (1978) has identified several types and sources of test bias. Several of Flaugher's categories are described below.

Differences in the mean test scores of two populations have been interpreted as an indication that the test in question is biased in favor of the high-scoring group. This interpretation has been offered for the relatively low scores obtained by many American minority-group children on standardized tests of intelligence (Williams, 1971), based on the rationale that such children are seldom exposed to the content found in most intelligence tests and are therefore unable to apply their intellectual abilities to the tasks. Other authors have argued that such group differences represent real differences in intellectual ability (Jensen, 1980; Sandoval, 1979). In fact, such differences may indicate neither test bias nor inherent group differences. They may reflect differences in educational opportunity which have resulted in real group differences in educational achievement (Flaugher, 1978; Rosenbach & Mowder, 1981). Flaugher suggested that much of the published debate regarding bias in psychological testing reflects confusion regarding the tests' status as aptitude or achievement measures. If interpreted in the former manner, low scores on such tests may discourage extensive remediation of academic difficulties. In the latter case, low scores would be interpreted as indicators of the need of special education or reform of the educational system.

Test bias may take the form of differential validity, where the test's accuracy in predicting a criterion varies across groups (Flaugher, 1978). Differential validity is operationally defined by statistically significant differences between the slopes, intercepts, or standard errors of estimate from the respective regression equations of the groups being compared (Cleary et al, 1975; Jensen, 1980). Humphreys (1973) noted that this set of comparisons should be conducted with samples of equal size, rather than comparing a minority group sample's statistics to those of the standardization sample for the majority group. The smaller minority group sample would be required to have a larger test-criterion correlation than that of the standardization group for significant departure from 0.

The problem of test bias through selection of a criterion measure is related to the problem of differential validity. An unreliable criterion may appear to be underpredicted for a minority group, i.e., group mean differences are smaller on the criterion than on the predictor being validated, as an artifact of the low reliability of the former (Flaugher, 1978). Earlier in the present report, it was noted that prediction of academic success from test scores assumed continuity in the form and quality of schooling from the time of testing on the predictor to the time of measurement of the criterion score. The validity of school achievement as a criterion is a confounding issue in some attempts to assess cognitive skills with less bias against minority groups. These attempts are described later in this report.

Flaugher (1978) noted that test bias is sometimes defined as qualities of the test items or tester characteristics which might differentially affect children within certain minority groups by highlighting their minority status and affecting their rapport with the tester. The common theme in this and other definitions of test bias offered by Flaugher is that properties of the test or testing situation affect members of one population in a manner which confounds the inferences made from the test results.

As described above, the assessment of test bias may be confounded by differences in the manner in which members of various populations interact with the instruments used to measure test validity. The examination of test bias is also confounded by ambiguities in the definition of population boundaries. Chapter II.C of this report describes several examples of studies in which the designation of culture (e.g., Native vs. nonNative culture) as an explanatory variable is uninformative and often misleading. Some procedures employed to reduce bias in testing are also linked to dubious assumptions regarding the definition of cultures and culture membership.

Methods of Reducing Test Bias

Methods of reducing the likelihood and damage of biased test interpretation have been addressed to properties of both the test and the testing procedure. Dillon and Stevenson-Hicks (1983) have described a series of procedures for integrating standardized testing procedures with examiner prompts to encourage elaborated item responses from the child. Gitmez (1972) examined the effect of providing or eliciting information on item difficulty to children, finding an interaction between socioeconomic status and the relative effectiveness of these two procedures.

The derivation of test norms for a specialized population has been advocated as a means of reducing the risk of biased test use with members of that population. Elliott and Bretzing (1980) stated that comparison of local norms with those of a larger standardization group (e.g. national norms for the same age group) provide information on the relative achievement of the local group and an indication of possible bias in the test. The availability of local norms also allows a psychologist to examine a child's achievement in reference to both groups.

The development of the System of Multicultural Pluralistic Assessment (SOMPA) (Mercer, 1979) was based on the rationale that local norms allowed a better estimate of a child's ability to learn. Mercer derived norms for children in various ethnic and socioeconomic status groups from samples of children in California. The pluralistic assessment model assumes that group differences in mean scores on standardized tests reflect bias in the test or testing procedures (Mercer & Ysseldyke, 1977). Standardized tests such as the WISC-R are included in the SOMPA battery, along with measures of social behavior and sources of medical information. Mercer's model calls for the comparison of children's scores to the score distributions for the various populations to which they belong. Estimated Learning Potentials (ELPs) are derived from their achievement relative to their own normative groups.

The provision of local norms is an inadequate solution to the problem of bias in testing. Ricks (1981) notes that the use of local norms for educational planning or vocational counselling assumes that the child will remain a member of that group. Comparison of test scores to both national and local norms, as recommended by Elliott and Bretzing (1980) and Mercer (1979), appears to address this problem. However, the accuracy of a prediction of achievement relative to some normative group may be a factor of the interval of time between testing and migration to a new cultural or educational setting. The educational background of a given child may not reflect the norm group to which he or she is assigned. In summary, the use of local norms to allow assessment of children's abilities according to standards which are specific to their educational and cultural background requires justification for the operational definitions of culture which determine the group membership ascribed to them.

Oakland (1980) found that the SOMPA ELPs of black and Chicano children had lower correlations with school achievement measures than were obtained by WISC-R IQs. This finding illustrates the point that an index, such as the ELP, which is thought to be less influenced by variability in educational or cultural experience than the IQ, may be a poorer predictor of achievement in the regular school system. Jirsa (1983) noted that the use of ELP's was descriptive, not prescriptive. The calculation of an unbiased index of potential for learning begs the question of how an unbiased learning environment would operate, i.e., under what conditions ELP-predicted achievement scores would be realized. Jirsa argued that children should not be barred from needed programming because their performance can be classed as average for a particular cultural group.

Some authors have suggested that nonbiased assessment requires the incorporation of all the assessment models described at the beginning of this chapter, providing information on the child's medical history, language ability, the child's ability to function in the social setting of the school, and the effectiveness of the school for other children in its care (Brady, Manni, & Winikur, 1983; Reschly, 1979). The SOMPA assessment battery also employs a wide range of normative and criterion-referenced tests, plus interviews and behavior checklists, to place information about the child in an ecological context. Although the integration of these models and testing procedures provide additional information which might be expected to decrease the likelihood and impact of over--interpretation of a biased test, the validity of inferences regarding any one test remains an empirical question.

The use of local norms addresses the issue of test bias as group differences in mean scores or the probability of assignment to special education classrooms. These adjustments to test scores are designed to insure that the scores are fairly applied but do not insure that inferences from the test scores are actually generalizable across samples. Flaugher's (1978) argument that group differences need not reflect test bias were reviewed above. Conversely, the absence of such differences does not insure that specific test interpretations are equally valid across groups. Berry (1969) has argued that generalization of behavior interpretation across cultural settings requires evidence that the behavior in question has functional equivalence, as well as metric equivalence, across settings. Standardization of test scores for a new population to the scale applied to a comparison population may be construed as an attempt to insure metric equivalence. For example, WISC-R subtest scores based on NWT norms, scaled to means of 10 and standard deviations of 3 and approximately normally distributed, allow a psychologist to make the same inference about an NWT child's rank on the subtest, relative to other NWT children, that an identical score, based on U.S. norms, would imply about an American child's ranking on the subtest. Functionally equivalent behaviors are responses to situational demands, shared by the two settings, to achieve identical goals in the two settings. To extend the above example, if the cognitive operations leading to a correct response to a WISC-R item are not similar across settings, or if the relationship of ability in the operations to academic success is dissimilar across settings, the tests are not functionally equivalent and generalization of inferences regarding cognitive or academic abilities is not valid. Cross-cultural psychologists have proposed that comparative studies of cognition employ a combination of psychological experimentation with anthropological fieldwork to test and generate, respectively, hypotheses about the cognitive operations employed in completion of a given task in different settings (Cole & Bruner, 1971; Laboratory of Comparative Human Cognition, 1979).

Functional equivalence appears to correspond to generalizability of construct validity. Factor analysis has a role to play in the generalization of construct definitions, as well as in their derivation. Buss and Royce (1975) recommended the abandonment of cross-cultural comparisons at the level of observed variables in favor of analysis of scores on factors for which structural invariance across settings has been demonstrated. The multivariate research approach advocated by Buss and Royce is similar in principle to the construct validation research plan proposed by Messick (1972), which was discussed in the present report in regard to factor theories of intelligence. Given Messick's well-documented argument that factor structures change with maturation and practice on the individual tests, generalization of a factor interpretation of a test or test battery to a new population would require replication of the implicit factor structure and experimental evidence for that interpretation of the factor(s).

Summary and Conclusions

The introduction to Chapter II.A included the assertion that the validity of present applications of intelligence testing is limited by the evidence for the viability of the construct of intelligence implied by a given application and evidence that the test score reflects ability on that construct. Present applications of intelligence testing to educational programming imply the definition of intelligence factors as ability traits which influence performance on a series of specific tasks. Remedial education is therefore directed to improving skills on relative cognitive weaknesses among the ability factors or to training children to apply their cognitive strengths to new tasks.

The nature and pervasiveness of the underlying factors has been a matter of debate among intelligence theorists, spawning a large array of factor analytic methods and test batteries to identify and measure, respectively, constructs as abstract and global as general intelligence and as specific as memory for figural relations. Although factor analysis is a useful method for identifying ability clusters or traits, it is not capable of explaining the nature of the factor, i.e., whether the factor reflects shared cognitive processing demands, capacity requirements (such as memory-store capacity), task stimuli properties, or some combination of these characteristics. Multivariate experimental research has been suggested as an aid to the exploration and validation of ability factors which intelligence tests are believed to measure.

The factors which influence a test score may change with changes in the examinees' practice with the task, cognitive and/or physical maturation, level of formal schooling, etc. Consequently, generalization of factor interpretations of a test to a new population requires evidence that the expected factors may be identified by factor analysis of test scores of members of that population. Generalization of inferences regarding the nature of the factors would require further experimental and correlational research. The generalizability of clinically--applied factor interpretations of WISC-R scores to arctic children is therefore dependant upon replication of the factor patterns which support those interpretations. Chapter II.B describes the factor models currently applied to the WISC-R and examines the research pertaining to their reliability and validity for nonarctic populations.

B. Interpretation of the WISC-R

Various definitions of intelligence and theories regarding its expression were described in Chapter II.A. These theories are reflected in the test batteries developed by their authors. Wechsler (1974) claimed that the WISC-R is not based on any particular definition of intelligence, other than as a global or composite entity which may be inferred from the expression of an assortment of abilities under a variety of conditions. Ultimately, "intelligence is the overall capacity of an individual to understand and cope with the world around him" (1974, p. 5). Wechsler argued that this definition differs from other theoretical definitions by emphasizing the multifaceted nature of intelligence and avoiding the isolation of any one ability as crucial to the definition. He specifically stated that general intelligence is not equated with intellectual ability.

The strength of the WISC-R as an instrument for psychological assessment is in part a function of the number and variety of its subtests (Wechsler, 1974). This section of Chapter II discusses the ways in which subtest and IQ profiles are interpreted to assess the various facets of intelligence expressed by an individual child. First, the organization of the subtests into the Verbal and Performance

Scales is described, followed by a brief description of current methods of profile analysis. The majority of this section deals with factor analytic studies of the WISC-R subtests with various populations and with attempts to ascribe psychological meaning to the resulting factors. The similarity of WISC-R factor patterns to those of the 1949 WISC is evaluated to determine whether the research findings compiled on the latter test may be generalized to the former to aid factor interpretation. Factor interpretation is then explored through discussion of the relationships of WISC and WISC-R subtests to tests and other variables extraneous to those batteries. The implications of the factor analytic results for the various interpretive models is then discussed. Finally, the generalizability of the interpretive models for the WISC-R to several populations defined in demographic terms, as opposed to clinical populations, is examined.

Organization of the Scales

The 12 subtests of the WISC-R are revised editions of the twelve subtests which comprised the original WISC (Wechsler, 1949). The WISC-R was standardized on 200 children of both sexes at each of 11 age levels between 6½ years and 16½ years (Wechsler, 1974). The subtests were scaled to means of approximately 10.0 and standard deviations of approximately 3.0 for each age level. As in the WISC, the Verbal and Performance Scales are comprised of six subtests each. The Verbal Scale subtests are as follows:

1. Information (Inf.) requires the child to provide factual answers to questions regarding historical dates and personalities, geography, etc.
2. Similarities (Sim.) requires the child to explain how two items, such as two specific parts of the body, are alike. Higher scores are awarded for identification of a shared conceptual category, such as "sense organs", than for identification of common features, such as "both part of the face".
3. Arithmetic (Ari.) involves counting, mental calculation, and problem solving.
4. Vocabulary (Voc.) requires the child to provide definitions for words presented orally by the examiner.
5. Comprehension (Com.) requires the child to provide solutions to practical and social problems and explain the purpose for certain social institutions and practices.
6. Digit Span (D.S.) has two parts. In Digit Span Forward (DSF) the examiner orally presents a string of numbers to the child, who must immediately repeat them from memory without error. After the successful repetition of at least one of two trials with digit strings of a given span, or length, a new set of strings is presented, with the span length incremented by 1. Digit Span Backward (DSB) requires the child to repeat each digit string in reverse order.

The Performance Scale subtests are as follows:

7. Picture Completion (P.C.) requires the child to identify an essential item which is missing from a line drawing.
8. Picture Arrangement (P.A.) involves the resorting of a series of line drawings so that their order depicts a story.
9. Block Design (B.D.) requires the child to reproduce pictorial designs with a set of colored blocks.
10. Object Assembly (O.A.) is essentially a jig-saw puzzle, comprised of four figures which the child must assemble.
11. Coding (Cod.) involves a test sheet with rows of empty squares, each underneath a square with a digit. Using a key which depicts the symbol associated with each digit, the child must fill in as many squares as possible, with the appropriate symbols, in a limited amount of time.
12. Mazes (Maz.) requires the child to trace an escape route, with a pencil, through a series of mazes drawn on paper.

The sum of scaled scores on Inf., Sim., Ari., Voc., and Com. is converted to the Verbal Intelligence Quotient (VIQ), while the sum of scaled scores on P.C., P.A., B.D., O.A., and Cod. is converted to the Performance Intelligence Quotient (PIQ). The sum of all ten scaled scores is converted to the Full Scale Intelligence Quotient (FSIQ). Digit Span and Mazes are optional subtests. They may be administered but they are not used in the calculation of VIQ, PIQ, or FSIQ, unless one of the mandatory subtests is invalidated.

Profile Analysis and Interpretive Models

In his recent textbook on psychological assessment of children, Sattler (1982) described five methods of analyzing the profile of a child's WISC-R subtest scores and IQs. These methods were:

1. Comparison of the Verbal and Performance IQs;
2. Comparison of each of the Verbal Scale subtest scores to the mean of those scaled scores;
3. Comparison of each of the Performance Scale subtest scores to the mean of those scaled scores;
4. Comparison of each subtest score to the mean of all subtest scores;
5. Comparing sets of individual subtest scores to each other.

The purpose of profile analysis is to identify the child's cognitive strengths and weaknesses. Wechsler (1974) has tabled the size of VIQ-PIQ discrepancies required for statistical significance, at the 5% and 15% level, for each age level for which the test has been normed. He has also provided the corresponding minimum discrepancy required for differences between subtest scaled scores. VIQ-PIQ or subtest discrepancies which exceed the tabled values may be considered reliable, i.e., the differences are too large to be attributed to random measurement error.

The range of scaled scores over all subtests, or subtest scatter, has also been considered a clinical indicator of psychological impairment (Kaufman, 1976a; Matarazzo, 1972). Kaufman (1976a) provided tables of norms for scatter indices, based on the distribution of the index in the standardization sample and suggested that judgements on the clinical significance of scatter indices be based on the frequency with which indices of that size appear in the normal population.

Bannatyne (1968) devised a procedure for analyzing the profile of WISC scores which was intended, with the aid of other perceptual and linguistic test batteries, to aid in the diagnosis of learning disabilities (handicaps to achievement in a specific academic skill, such as reading or mathematics, which is not attributable to low general intelligence). He categorized WISC subtests into the categories of Conceptualization (Con), Spatial (Sp), and Sequential (Seq) ability. The composition of these categories was revised (Bannatyne, 1974) to accomodate suggestions proposed by Rugel (1974b) in a review of factor analytic research of the WISC. A new category, Acquired Knowledge (AK), was added in the recategorization of subtests. The revised categories, which are applied in profile analysis of the WISC-R, are as follows:

1. The Conceptualization category score is calculated as the average of the scaled scores for Similarities, Vocabulary, and Comprehension;

2. The Spatial category score is the average of Picture Completion, Block Design, and Object Assembly scores.
3. The Sequencing category score is the average of Arithmetic, Digit Span, and Coding;
4. The Acquired Knowledge category score is the average of Information, Arithmetic, and Vocabulary (Bannatyne, 1974).

The existence of four factors in operation among the subtests was supported by the results of the factor analysis of subtest intercorrelations for the WISC standardization sample (Cohen, 1959). Those factor analytic results are briefly reviewed later in this section of Chapter II, in an attempt to assess the applicability of WISC-based interpretive procedures and models to the WISC-R.

The purpose of the present study was defined in Chapter I as an examination of the validity of three interpretive models of the WISC-R for educational assessment in Canada's eastern arctic. The first of these models, the Verbal-Performance model, was built into the organization of subtests into Verbal and Performance Scales and is implicit in the interpretation and comparison of VIQs and PIQs. Bannatyne's model was based on factor analyses of the WISC and clinical objectives for a specific population of children. The third model, Kaufman's three-factor model, was based on factor analyses of the subtest intercorrelations for the WISC-R standardization sample. Kaufman's model will be described in the context of a review of factor analytic studies of the WISC-R, which is presented immediately below.

Factor Analyses of the WISC-R

Many of the procedures for clinical interpretation of WISC-R subtest scores, which were discussed under the topic of profile analysis, are based on rationalizations of the task demands of the individual subtests, factor analyses of the 1949 WISC, correlations of the subtests or similar tasks with measures not included in the WISC-R. Another major source of interpretive guidelines has been the series of factor analyses of the subtest intercorrelations for the eleven age groups included in the U. S. standardization of the test. Subsequent analyses have been conducted with various samples from special clinical or demographic populations in attempts to replicate and generalize the factor structure found for the standardization samples. The results of these factor analyses are described and discussed below, followed by a discussion of attempts to interpret those factors.

Analysis of the Standardization Sample

The subtest intercorrelations obtained for each of the eleven age groups in the U. S. norming sample have been factor analyzed by several methods with similar, although not perfectly consistent, results. Among the first of these to be published was the analysis by Kaufman (1975), who was involved in the standardization of the test. The results of this analysis are described in detail, while discussion of other studies will be limited to differences in method and/or results from those of Kaufman.

Kaufman's (1975) analysis began with the extraction and rotation of principal components for each of the age groups. The number of components with eigenvalues greater than 1.0 provided an initial guide to the number of factors to rotate. The results of this method suggested two factors for six of the age groups, while three factors were suggested for the remaining five groups. However, the value of the third eigenvalue ranged from 0.9 to 1.1 across all eleven groups, which Kaufman appears to offer as evidence for the significance of the third factor for each group. He then performed principal axis factor analysis on each set of data, with squared multiple correlations for the variables as initial estimates of their communalities. Orthogonal (varimax) and oblique (oblimax and biquartim) rotations of the resulting factor matrices were performed for up to five factors. Kaufman claimed that the final criteria for determining the meaningfulness of a factor was: the presence of at least one loading greater than .20; the appearance of the factor in the results of at least six age groups; and correspondence to developmental theory.

The orthogonal rotation of two factors resulted in a factor pattern which resembled the Verbal and Performance Scale organization of the subtests. Kaufman (1975) accepted an orthogonal three-factor solution as the most reasonable across all but two age groups.

1. The first of these factors was labelled Verbal Comprehension (VC) and was defined by high loadings for Information, Similarities, Vocabulary, and Comprehension. The reader should note that the latter three subtests formed Bannatyne's Conceptualization (1974) category.
2. Factor II, labelled Perceptual Organization (PO), was defined by Picture Completion, Picture Arrangement, Block Design, Object Assembly, and Mazes. Bannatyne's Spatial category is subsumed in this factor.
3. The third factor, comprised of Arithmetic, Digit Span, and Coding, was labelled Freedom from Distractibility (FD). This factor is identical to Bannatyne's Sequencing category.

Rotation of four factors was required to obtain the FD factor for the 6½ and 14½ year age groups. Kaufman argued for a four-factor solution for these groups but claimed that the extra factor, which contained Performance Scale subtests, was uninterpretable. He expressed satisfaction that each subtest was a good measure of a single factor.

Kaufman's (1975) assertion that the factor patterns for the eleven age groups represent reliable, consistent models of simple structure is somewhat surprising in the light of examination of his tables of factor loadings and his discussion of several discrepancies among patterns for various groups. For example, Picture Arrangement had VC loadings greater than .3 for seven ages, and its VC loading exceeded its PO loading for three of these groups. Picture Completion had VC loadings above .3 for eight age groups, although its largest loading was on the PO factor. Coding had large FD loadings for only five groups, while Information had large FD loadings for six ages. In fact, the orthogonal factor patterns for the eleven age groups, as presented and accepted by Kaufman, exhibit neither the consistency nor the simple structure he suggests.

The results of oblique rotation of the factor matrices were similar to those of the varimax rotation (Kaufman, 1975), with two important exceptions. The third factor was more decisively defined by Arithmetic, Digit Span and Coding, with lower Information loadings than were obtained by orthogonal rotation. The VC factor for the 6½ and 7½ year old samples was much like a general factor, while the other factors contributed less to the variance. Presumably this latter result of oblique rotation is what led Kaufman to favor the varimax results, although the reasons for that choice are not made explicit. The present author disagrees with Kaufman's argument that the varimax results are the most reasonable and interpretable. Wechsler's (1974) definition of global intelligence would suggest that mental factors are correlated rather than uncorrelated. The stonger presence of the general factor at younger ages may reflect adherence to Garrett's (1946) hypothesis that intellectual abilities become increasingly differentiated throughout childhood. Principal component analysis of both the 6½ and

7½ year data, as well as that of the 9½ year sample, returned only two eigenvalues greater than 1.0 (Kaufman, 1975; Silverstein, 1977). However, a clear developmental trend toward more factors is not evident among the older samples. These theoretical considerations, paired with the improved adherence to simple structure, are the basis for the present author's argument that oblique rotation gives a clearer understanding of the factors underlying the tests. Factor analytic studies of the standardization sample by other methods of extraction and/or rotation are described below.

Silverstein (1977) extracted principal axis factors by the maxplane method (Eber, 1966). The two-factor solution reflected the organization of the Verbal and Performance Scales, with the exception of Coding, which did not load on either factor. Silverstein calculated congruence coefficients for the comparison of factor patterns between pairs of age groups. (This coefficient is a measure of the similarity of loadings for two factors, interpreted in much the same manner as a correlation coefficient. The congruence coefficient is explained in detail in Chapter II.D.) The median congruence coefficients for the VC and PO factors were very high at .89 and .91, respectively, while the median coefficient for the FD factor was a more modest .75. Silverstein concluded that the two- and three-factor solutions were both relatively stable and that neither could be said to be right or wrong.

Hierarchical factor analysis resulted in only two primary factors, with the general factor accounting for an average of 36% of the common variance for the eleven age groups (Wallbrown, Blaha, Wallbrown, & Engin, 1975). Wallbrown et al. interpreted the primary factors as Vernon's (1965) *v:ed* and *k:m* factors, comprised of the Verbal and Performance (minus Coding) subtests, respectively.

The validity of the Verbal and Performance Scales as two factors was tested directly by multiple group factor analysis of the data of the 7½, 10½, and 13½ year standardization samples (Ramanaiah, O'Donnell, & Ribich, 1976). (This method of analysis is described in Chapter II.D.) Ramanaiah et al. concluded that the two-factor model was supported. This support was reinforced by Silverstein's (1982) multiple group analysis of both the two-factor model and the three-factor model proposed by Kaufman (1975). Silverstein again concluded that the data was equally supportive of either of the two models, although he suggested that two factors might be more parsimonious.

Silverstein and Legutki (1982) compared the results of principal factor analysis, multiple group analysis, and hierarchical factor analysis of both the WISC and WISC-R data for the same three age groups and concluded that, with the exception of the large influence of *g* on the results of the hierarchical analysis, that the various methods resulted in similar conclusions. This summary appears to ignore the absence of the FD factor from the hierarchical findings.

Cluster analysis of the twelve subtests supported a three-factor solution for the matrix of correlations averaged across age groups (Silverstein, 1980). The results for individual age groups provided some support of the two-factor solution as well, as pentads resembling the two IQ scales appeared for some samples. Silverstein concluded that these results simply confirmed the equivocal nature of the results of factor analysis of the WISC-R and stated that the ultimate test of either factor model would be its clinical utility.

To provide further insight into the importance of common factors to subtest scores, Silverstein (1976) divided each subtest's variance into common variance, subtest-specific variance, and error variance. Common variance was expressed as the squared multiple correlation (SMC) of a given subtest with the other subtests. specific variance was calculated as the difference between the subtest's reliability coefficient and its common variance. Error variance was calculated as 1.0 minus the reliability coefficient. The VC subtests and Block Design tended to have the most common variance (52%-65%), while the remaining subtests had relatively more specific and error variance. Digit Span, Picture Arrangement, Coding and Mazes all had more specific variance than common variance and the latter two had more error variance than common variance. The ranking of subtests by contribution of common variance above corresponds closely to their ranking by the relative contribution of the general factor, whether *g* loadings are calculated by hierarchical analysis (Wallbrown et al., 1975) or as the loadings on the first unrotated principal component (Kaufman, 1975). The relative contributions of common, specific, and error variance have implications for the validity of the interpretation of individual subtest scores. These implications are discussed later in this chapter. The generalizability of the factor patterns derived from the standardization samples' scores to those of other populations is examined first, as confidence in the application of factor interpretations of the WISC-R requires identification of the populations whose scores are defined by those factor models. Some of the studies reviewed for this purpose have focused on subsamples within the U.S. standardization sample.

Generalizability of WISC-R Factor Patterns

The issue of stability of factor patterns across age groups was raised above in regard to the standardization samples. The two-factor solution appeared to be stable across groups, while evidence for the stability of the third factor was less conclusive. Garrett's (1946) theory, that mental abilities differentiate with increasing age throughout childhood, was cited as grounds for expecting an increase in the number of factors for the older groups. Reynolds and Gutkin (1980) grouped the standardization sample scores for ages 6½ to 10½ and for ages 11½ to 16½ and factor analyzed the subtest correlations for each of these two pooled groups. The congruence coefficients for three corresponding factors were all above .95. This finding suggests that the definition of these factors is not related to age. However, the reader is reminded that decisions regarding the saliency of some test loadings on a given factor, such as those of Picture Arrangement on the VC factor or Information on FD, were not consistent across the eleven age groups.

Populations Defined by General Intelligence

Evidence regarding the generalizability of the two- and three-factor clinical models for populations of mentally retarded (MR) children is inconsistent, particularly in relation to the FD factor. Schooler, Beebe, and Koepke's (1978) principal components analysis of the scores of 127 educable mentally retarded (EMR) children resulted in the extraction of only two components, which were similar in definition to the Verbal and Performance Scales. However, the analysis excluded Digit Span and Mazes. The absence of the former subtest precluded the possibility of identifying the FD factor as defined by Kaufman (1975). Cummins and Das (1980) extracted three principal components, closely corresponding to the VC, PO and FD factors, from the WISC-R subtest scores for 95 adolescents with IQs ranging from 55 to 80. Van Hagen and Kaufman (1975) analyzed the scores of 80 children of both sexes from special education classrooms and state institutions. These children ranged in age from roughly 3 years to 16 years; in WISC-R Full Scale IQ, from

40 to 79, with a mean FSIQ of 50.6. Principal components and factor analysis resulted in factors resembling Kaufman's VC and PO factors, but the FD factor only emerged if four factors were rotated. Comparison to the factor pattern of a small sample of average-IQ children revealed that the congruence coefficients for noncorresponding factors ranged from .58 to .87. In fact, the "FD" factor for the MR sample was more like the nonretarded sample's "VC" factor than the latter group's "FD" factor. The most striking difference between the results for the retarded sample of Van Hagen and Kaufman and those of the standardization sample is the ranking of the subtests on *g* loadings. Picture Arrangement and Picture Completion had the highest loadings on the first unrotated principal factor, while Coding had the fourth largest such loading. In contrast, the first two of these were considered only fair measures of *g* according to analysis of the U.S. standardization sample and Coding was labelled as a poor measure of *g* in the same reports (Kaufman, 1975, 1979a).

Hierarchical factor analysis of the scores of 79 MR children resulted in three primary factors and a general factor under which two of the primary factors were nested (Vance, Wallbrown, & Fremont, 1978). The first primary was labelled *v:ed* and was defined by the four subtests in Kaufman's VC factor (Inf., Sim., Voc., and Com.). The second primary was labelled *k:m*, with large loadings for Picture Completion, Picture Arrangement, and Object Assembly and more moderate loadings for Block Design and Cod. The third and isolated factor, labelled Stimulus Trace (ST) or short-term memory was defined by Arithmetic, Block Design, and Digit Span. Vance et al. concluded that the factor structure for the MR sample was more complex than the structure derived by hierarchical analysis of the standardization sample (Wallbrown et al., 1975). This result is similar to the findings for hierarchical and multiple group factor analysis of WISC scores of MR samples by Baumeister and Bartlett (1962a, 1962b). In one of these studies (Baumeister & Bartlett, 1962b), the ST factor corresponds exactly to Kaufman's FD factor, i.e., the salient loadings were for Arithmetic, Digit Span, and Coding. Baumeister and Bartlett suggest that the variance of short-term memory is too restricted among children of average general intelligence for the factor to be defined through hierarchical analysis of the standardization samples' data. This explanation is speculative, but verifiable. However, the fact that a third principal axis factor was significant for the upper ages of the standardization sample (Kaufman, 1975) discredits the theory that the factor measures skill in a process which is too simple to be reflected in the correlations of average-ability children.

Groff and Hubble (1982) factor analyzed the data for young (average age of 10 years) and older (15 years) retarded children and concluded that the FD factor was more prevalent for the younger sample. This conclusion was based on the percentage of common variance attributed to the factor, but the pattern of subtest loadings is also discrepant across age groups.

The factor pattern obtained for gifted children is also partly discrepant from that obtained for the full range of IQs. Factor Analysis of the subtest scores for 946 children whose FSIQ exceeded 120 resulted in three factors, which the authors labelled VC, PO, and FD (Karnes & Brown, 1980). However, comparison to studies with the standardization sample are confounded by the fact that Digit Span and Mazes were excluded from the analysis. The results may also be affected by restriction of range of the scores, as the standard deviations of the VIQ, PIQ, and FSIQ are 8.99, 10.88, and 7.14, respectively, in contrast to the standard deviations of approximately 15.0 for the standardization sample. The "FD" factor is defined by Arithmetic and Picture Completion and interpreted as freedom from distractibility in tasks involving part-whole relationships. The reader may be excused for suspecting that this label is an elaborate hedging of bets for interpretation of a factor with only two salient loadings.

The above studies did not provide support for the generalization of the Verbal-Performance or Kaufman factor models to populations of retarded children. However, the nature of discrepancies from those models was not consistent across studies. The factor pattern obtained for gifted children may be affected by restricted variance and the exclusion of subtests. The importance of these findings is open to question. Since the scores of these two groups of children fall on the extremes of the IQ dimension by definition, the correlations of subtest scores within the groups might reasonably be expected to be depressed. If the reference score for comparison of a child's mental ability is the average score for that age group, it seems illogical to define and combine subtests in a different manner for children of a given age but with varying FSIQ scores. To do so would make it impossible to define factors with sufficient range to discriminate among children of different ability levels. The remaining studies of the generalizability of factor models concern target populations which are defined by criteria in addition to, or in exclusion of, their average Full Scale IQ. However, the degree to which the subtest correlations for a sample, and hence the factor pattern, are affected by restriction of range of the scores limits the validity of assertions that a given subtest measures some ability other than that which it measures for the standardization sample.

Clinical Samples

A number of studies have attempted to determine the WISC-R factor structure for clinical populations, i.e., populations of children who have been diagnosed as learning disabled, emotionally disturbed, etc., after referral to a psychologist for assessment. Schooler et al. (1978) found two principal components for a sample of children diagnosed as learning disabled, but the exclusion of Digit Span from the analysis may be responsible for the failure to identify a third factor. Petersen and Hart (1979) divided a clinic sample of teacher-referred children into three subgroups on the basis of diagnoses of learning disabilities (LD), emotional disturbances (ED), or the absence of a specific handicap (NH). The subtest scores of these groups were factor analyzed separately and as a single sample. Although three factors were extracted for all samples, the definition of the third sample is widely discrepant across groups. None of these closely resembles Kaufman's FD factor. Much stronger replication of the Kaufman three-factor model was found for samples of children referred for learning disabilities and behavior problems (Stedman, Lawlis, Cortner, & Achterberg, 1978), male juvenile delinquents (Hubble & Groff, 1981), and a sample of psychiatric patients, aged 6 years to 16 years, which excluded children diagnosed as MR or LD (Hodges, 1982). One study replicated the Verbal-Performance two-factor model for a sample of children referred for school behavior problems (Finch, Kendall, Spirito, Entin, Montgomery, and Schwartz, 1979). However, these researchers included the children's VIQs and PIQs as separate tests in the analysis. Since ten of the subtests, including Arithmetic and Coding from the FD set of tests, contribute equally to the IQ calculated for their respective scale, this procedure might be expected to bias the factor pattern toward a solution centered around the IQs.

Profile and factor analysis studies with clinical groups are often difficult to interpret due to characterististics of the data collection procedures. The samples may be heterogenous in regards to diagnostic category, age, or other variables. This situation makes it impossible to identify the population represented in the sample, unless the reader is simply concerned about generalizing assessment principles to all teacher-referred children in the particular school district or clinic catchment area studied. As in the Schooler et al. (1978) study, the analysis of scores obtained for clinical purposes may be affected by the absence of subtest data which was not required for those purposes.

The emphasis thus far has been on the stability and generalizability of the Verbal-Performance and Kaufman factor models across methods and populations. Even where the factor patterns have been stable across samples, the interpretation of the factors represented has varied across authors. Factor interpretation is the emphasis of the following sections of Chapter II.B, beginning with an examination of the validity of generalizing the results of WISC studies to interpretation of the WISC-R.

Comparison with the WISC

Attempts to explain the factors underlying the WISC-R subtests have included experimental manipulation of the task demands, factor analysis of larger batteries of tests with WISC-R subtests included, and examination of the literature pertaining to similar tasks. Some of these experimental and correlational studies were conducted on WISC subtests. The organization of the subtests into Verbal and Performance scales and the use of Bannatyne's (1974) subtest recategorization reflect a transfer of models for clinical inference from the WISC to the WISC-R. The comparability of the factor structures for these two tests, and hence the validity of this transfer, is explored below.

The most widely-cited factor analysis of the WISC subtest correlations was conducted on the data of the 7½, 10½, and 13½ year age samples for the U.S. standardization of that test (Cohen, 1959). Cohen extracted five centroid factors for each age group and independently rotated these to oblique simple structure. Second-order factor analysis of the correlations among factors was conducted to determine the *g* loadings for the subtests. The five factors were inconsistent in definition across age groups, if identical decisions on the saliency of a given test's loading on a given factor is the criteria for consistency. The factors were labelled consistently, however, as Verbal Conceptual I and II, Perceptual Organization, Freedom from Distractibility, and Coding. The reader may note that the labels for the first four factors were adopted for Kaufman's (1975) WISC-R factors. The nature of these factors is discussed below.

The first Verbal Conceptual factor, (VC I), had large loadings for Information and Similarities at all ages, with Vocabulary, Arithmetic, and Comprehension loading at one or two ages. VC II had large Comprehension and Picture Completion loadings at all ages, with Vocabulary loading on this factor at the two younger ages; Similarities, at age 13½. Cohen (1959) suggested that VC I reflected school learning while VC II reflected judgement involving verbal manipulations (of symbols, ideas, etc.). The Perceptual Organization (PO) factor included Block Design and Object Assembly at all ages, Picture Completion and Mazes, each at two ages, and Picture Arrangement at one age. The Freedom from Distractibility (FD) factor had large Digit Span loadings at all ages, with large loadings for each of Arithmetic, Picture Arrangement, Object Assembly, and Mazes for at least one age group. Coding appeared as a separate factor, with a large Picture Arrangement loading at the upper ages. The stability of the factors across ages is not impressive, but these results have nonetheless contributed to the formulation and acceptance of Bannatyne's categorization for both the WISC and WISC-R. Bannatyne's (1968) original categorization had placed Picture Arrangement and Vocabulary in the Sequencing category in place of Arithmetic and Coding.

Direct comparison of the WISC and WISC-R factor structures has generally been confined to two- and three-factor solutions to determine the generalizability of the Verbal-Performance Scale organization and Kaufman's (1975) three factors. Swerdlik and Schweitzer (1978) administered the two tests in counterbalanced order to a clinical sample. Two- and three-component solutions were independently rotated to the varimax criterion. The resulting factor patterns were compared by calculating coefficients of concordance, which are conceptually similar to congruence coefficients. The two-factor solutions for the WISC and WISC-R were very similar, with concordance coefficients of .99 and .98. The third factor was less similar across tests, with a concordance coefficient of .77. Although the third-factor coefficient appears respectable for a sample size of 164 (if the statistic is considered as conceptually related to a correlation coefficient), the reader should note that the third factors for the two tests shared only one common salient subtest. This subtest was Picture Arrangement, which was not included in Kaufman's (1975) definition of the third factor. Therefore, the generalizability of this test comparison is limited. Even Swerdlik and Schweitzer's claim of generalizability to children referred to a school psychologist is open to question in the absence of a demonstration that their sample is representative of all such clinical samples.

The following studies were cited in regard to alternative methodologies for factor analyzing the standardization data of the WISC-R. Ramanaiah et al.'s multiple group application of the Verbal-Performance model to the WISC standardization data provided support for the two-factor organization of the WISC. Silverstein's (1982) multiple group analysis supported both the two- and three--factor solutions for the WISC, as it had supported these models for the WISC-R in the same study. Silverstein and Legutki (1982) reported that congruence coefficients for corresponding factors were as large for test-wise comparisons as for age group comparisons employing the same test.

The above studies, employing several indices of the similarity of factor patterns across samples and tests, have generally suggested that the factor structure of the WISC and WISC-R are very similar. However, comparisons based on the identification of loadings as salient or nonsalient have led to different conclusions. Whereas Cohen's (1959) analysis of the WISC led him to identify two verbal conceptual factors, Kaufman's (1975) WISC-R analysis analysis indicated that rotation of four factors led to a splitting of the Performance Scale factors. When only two or three factors are rotated, the above studies suggest that the VC and PO factors are fairly similar across tests; the FD factor much less so. There is no evidence in the above studies to support a four-factor interpretation of the WISC-R which is based on the factor analytic results of the WISC. This conclusion invalidates the calculation and interpretation of Bannatyne's (1974) Acquired Knowledge factor for WISC-R interpretation.

It is interesting to note that subtests in the WISC-R FD factor, for which the association with the WISC factor structure is weakest, were copied from the WISC with very few item modifications. Of 18 items in the Arithmetic subtest, 8 are taken directly from the WISC, while 5 items are modified WISC items. Digit Span and Coding items were all taken directly from the WISC, with the addition of color to Coding A as the only change. While Picture Completion has 11 new items among its total of 26, and Picture Arrangement includes only 3 unmodified WISC items among its total of 12, the remaining PO subtests were incorporated from the WISC with few changes. However, the numbers of new items in the VC subtests range from .30 to .47 of the total number of items in that subtest (Wechsler, 1974). The shift of Arithmetic into the FD factor, and Picture Arrangement into the PO factor, may reflect the effect of large-scale modifications to those subtests on the subtest intercorrelations.

The difficulties associated with the means of factor pattern comparison applied in the above studies are discussed in detail in Chapter II.D. These methodological concerns are also relevant for studies which have compared the factor structure of the WISC-R across populations defined by sex, socioeconomic status (SES), ethnic group or nationality. This caution is offered to temper acceptance of the results of such studies, which are discussed at the end of Chapter II.B and in Chapter II.C, until the reader has considered the methodological considerations discussed in Chapter II.D.

Interpretation of the Factors

Kaufman (1979a) has argued that interpretation of WISC-R scores should incorporate psychological theory regarding cognitive development. Research investigating the validity of specific theoretical interpretations of subtests or sets of subtests is described below. The review begins with a test of one theory through factor analysis of a subset of the 12 subtests and then examines the relationship of various subtests to extraneous variables. It will focus almost exclusively on the subtests in the Perceptual Organization and Freedom from Distractibility factors defined by Kaufman (1975). Of the former subtests, those in Bannatyne's (1974) Spatial factor, i.e., Picture Completion, Block Design, and Object Assembly, receive the most attention. The Spatial subtests, Digit Span, and Coding were taken from the WISC with minor or no modifications. Since much of the research on relationships of individual subtests to other variables has been conducted with the WISC version of the subtest, more confidence can be placed in such results for the subtests mentioned than would be possible for the VC subtests, Picture Arrangement or Arithmetic.

The emphasis on the PO and FD factors also reflects the greater interpretive ambiguity associated with the subtests in these factors. Factors composed of the Verbal Scale subtests have been labelled Verbal Comprehension (Kaufman, 1975), Conceptualization (Bannatyne, 1974), and Verbal:Educational (Wallbrown et al., 1975). All these terms seem to reflect an interpretation of the factor as a measure of academic knowledge and ability, or expresssion of intelligence through a verbal medium. Wechsler (1974) stated that the Verbal-Performance dichotomy was "primarily a way of identifying two principal modes by which human abilities express themselves" (p. 9). However, as the following review indicates, the relative importance of cognitive styles, brain disorders, and personality variables have been debated in regards to the PO and FD (or Sequential) factors.

Kaufman (1979a) suggested that the WISC-R subtests be organized according to the simultaneous-successive processing model of Das, Kirby, and Jarman (1979). Simultaneous processing involves multiple comparisons among stimuli or responses and the "construction of a spatial pattern or scheme" (Das et al., 1979, p. 52). Successive processing involves the temporal organization of stimuli and/or responses. Kaufman (1979a) had suggested that Similarities, Picture Completion, Block Design, and Object Assembly were tasks which required simultaneous processing, while Digit Span, Picture Arrangement, Coding, and Mazes required successive processing. Naglieri, Kamphaus, and Kaufman (1983) attempted to test the applicability of the processing model to the WISC-R subtests by reanalyzing the data from the original standardization sample (Wechsler, 1974), the MR sample of Van Hagen and Kaufman (1975), and the clinic-referred sample of Stedman et al. (1978). Naglieri et al. omitted Information, Arithmetic, Vocabulary, and Comprehension from the analysis claiming that these subtests were heavily influenced by school achievement and poor measures of the two processing modes. The remaining eight subtests were factor analyzed and two factors were rotated. The results did not support the hypothesized division into processing modes, as Picture Arrangement and Mazes both loaded on the factor labelled Simultaneous for two of the samples, while Similarities loaded on both factors for the standardization sample and the Successive factor for the other samples. Naglieri et al. concluded that the inconsistencies found, both across samples and with their hypotheses, reflected differences in task strategy which explained the relatively lower scores obtained by children such as those in the MR and clinical samples. For example, they suggested that the standardization sample used simultaneous processing to solve Similarities items, while the other samples applied only successive processing to that subtest. However, Matheson (in press) demonstrated that the Simultaneous loading obtained for Similarities with the Standardization data was a capitalization on factor score indeterminacy, as similar patterns could be obtained for each of Information, Vocabulary, or Comprehension, if analyzed in isolation of the other VC subtests. (Indeterminacy of factor scores is explained in Chapter II.D.) The implication of the inconsistencies and ambiguities in the Naglieri et al. results, Matheson's demonstration of the spurious nature of those results, and the discussion of factor score indeterminacy which appears later in this report, is that insight into the meaning of the three factors obtained by Kaufman (1975) is not to be gained by factor analysis of smaller subtests of the 12 subtests. Interpretation of the factors requires examination of their relationships to variables outside the WISC-R battery.

The factor analytic research reviewed thus far has been limited to analysis of the 12 subtests. The research reviewed in the following pages examines the relationships of the subtests to extraneous variables, as examined through factor analysis of larger batteries of tests, experimental manipulation of task demands, or simple correlations with criterion variables.

Perceptual Organization Subtests

The Spatial subtests, i.e., Picture Completion, Block Design, and Object Assembly have been theoretically linked to the cognitive style dimension labelled field independence/dependence. Field independence (FI) refers to the ability to perceive a figure as discrete from an organized ground, or to impose structure upon a field which has little inherent organization (Witkin, 1974). An inability to perceive and analyze a figure apart from the field is labelled field dependence. These poles of perceptual ability are thought to reflect analytic vs.

global modes of cognitive processing, or cognitive styles (Witkin, 1974; Witkin, Moore, Goodenough, & Cox, 1977). Field independence has been measured by such tests as the Embedded Figures Test (EFT), which requires the testee to find and trace a line drawing which is hidden within a complicated visual field. In the Rod-and-Frame Test (RFT), the testee is seated in a darkened room in which one wall supports a lighted rod, surrounded by a lighted frame, both of which are tilted at various angles throughout the test session. The subject must rotate the rod until it is perfectly vertical, utilizing kinesthetic cues while trying to ignore the disorienting cues of the tilted frame. These and similar tests are described in Witkin et al. (1977). Witkin (1974) has hypothesized that field independence is fostered by child-raising methods which encourage the child to explore and develop an identity separate from the parents. This theory has been examined in a great deal of cross-cultural research on cognitive processing, and the segments of that research which pertain to North American native populations are reviewed in Chapter II.C. The argument and evidence that the WISC-R Spatial factor measures field independence is examined below.

A series of field-independence measures loaded on a factor with the WISC Picture Completion, Block Design, Object Assembly and Mazes when centroid analysis was conducted on the test battery scores of a mixed-sex 11-12 year old sample (Goodenough & Karp, 1961). Arithmetic, Digit Span and Coding loaded with some paper-and-pencil measures of field independence on a factor labelled Attention-memory by Goodenough and Karp. Most of these results were replicated with a sample of 9-10 year old boys, but the reliability of the factors in the second study is limited by a low ratio of subjects to variables. Goodenough and Karp concluded that perceptual and intelligence tests shared a demand for the subject to overcome the context in which a critical stimulus is embedded.

Attempts to extend the relationship of the PO or Spatial subtests with field independence to older and younger groups is complicated by changes to both the Wechsler and FI tests for those age groups. Coates (1975) administered the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) and his own FI test, the Preschool Embedded Figures Test (PEFT) to a sample of children aged 4-5 years. Separate factor analyses for each sex resulted in a factor with the PEFT, WPPSI Geometric Design and WPPSI Block Design, which Coates labelled as Perceptual Analytic, and a factor labelled Verbal Comprehension. Coates rotated, but did not interpret, two additional factors for each sex. The fourth factor had large Picture Completion, Block Design, and Mazes loadings for the female sample, but incuded only Mazes for the male sample. This tendency for Mazes to be separate from the Spatial and FI tests in male samples, while joining those tests in female samples, was also found in the Goodenough and Karp (1961) study described above.

Karp (1963) factor analyzed a battery of tests which included Block Design, Object Assembly, Comprehension, Vocabulary, Arithmetic, and Digit Span from the Wechsler Adult Intelligence Scale (WAIS). Block Design and Object Assembly loaded on a pair of factors with other FI tests. These latter tests included the EFT and RFT as well as tests that required reformulation of a given problem to overcome initial assumptions that might be falsely inferred from the context of the problem. The second FI factor included those FI tests which were timed. As hypothesized, the FI factors did not include those tests which required the subject to ignore a field which was distracting but without disorienting cues. Karp concluded that the analytical ability reflected in FI tests was therefore distinct from both general intelligence and attention to the task.

Karp's (1963) conclusion that field independence is not simply a nonverbal facet of general intelligence was examined by Satterly (1979). Satterly attempted to partial general intelligence out of the relationship of EFT scores to several achievement and personality measures for 10-11 year old children by including a short paper-and--pencil IQ test in a principal components analysis. Although Satterly argued that the EFT provides predictive information on math and geography achievement in addition to the information provided by general intelligence scores, his methodology contained flaws which have been identified in

Humphreys and Parsons' (1976) critique of his earlier publications. The latter authors argued that general intelligence cannot be "partialled out" with a single IQ test, which is not a perfect measure of *g*. They conducted an hierarchical analysis of Satterly's original matrix and found that, while the EFT and achievement tests loaded together on the second-order factor, EFT loaded on a primary factor with other FI tests and the achievement tests loaded on a factor labelled Verbal Conceptual. Humphreys and Parsons concluded that the correlation between the EFT and school achievement is explained by general intelligence. Vernon (1972) conducted principal components analysis on several FI measures, personality, verbal and spatial tests, including Kohs Blocks (from which the WISC-R Block Design was adapted). Vernon reported a *v:ed* factor, a spatial--perceptual factor, and a factor labelled as *g*. Whereas FI tasks requiring physical manipulation of the stimulus, such as the RFT, formed a distinct spatial construct, Vernon concluded that the paper-and-pencil FI tests were simply measures of *g*.

The research reviewed above raises the question of whether the relationship between the WISC-R Spatial tests and FI measures reflects the influence of field independence on WISC-R scores or simply the influence of general intelligence upon both sets of tests. Swyter and Michael (1982) found that the Spatial subtests of the WISC-R loaded on a single factor when analyzed with a battery of perceptual tests, paper-and-pencil FI measures, and Piagetian conservation tasks. These authors concluded that the three WISC-R subtests and FI measures shared the influence of a common factor other than general intelligence, based on relatively low correlations of these tests with IQ scores found in the children's school records. Swyter and Michael suggest that this construct might be field independence, but also consider the possibility that the construct is simply skill with figural content.

Guilford (1980) interpreted field independence tests in the context of his Structure of Intellect theory, claiming that FI is a measure of flexibility, rather than cognitive analytic skills. Guilford suggested that the construct involved a willingness to make transformations (one of the product categories in his SI model), citing the difficulty which field dependent people experience in reformulating problems. He concludes that FI may be both a higher-order aptitude trait and an intellectual executive function, somewhat like Sternberg's (1981, 1982) metacomponents of intelligence.

Wachtel (1968) argued that the construct reflected ability to isolate and learn segments of a stimulus or task, based on his finding that field independent adults were superior to field dependent adults at identifying segments of designs which had been initially presented in their entirety. Although the field independent subjects in Wachtel's study had superior WAIS Vocabulary scores, partialling out this effect by analysis of covariance did not alter the conclusion that they were superior to field dependent subjects at learning segments of stimuli presented as more complex wholes. Wachtel (1968, 1972) concurred with Guilford's (1980) hypothesis that field independence is an ability and an executive function. Field independent people are able to analyze stimuli and problems in components, but are not restricted to that mode of analysis. This view is consistent with research demonstrating that while field independent subjects tend to be relatively insensitive to social cues and forget the names and faces of people they meet, unlike field dependent subjects, they are adept at global analysis and perception of social cues if they choose to attend to those aspects of a task or problem (Witkin et al., 1977)

Wachtel (1972) suggested that interpretation of tests such as Block Design, Object Assembly, and Picture Completion as measures of field independence must be conducted in the context of other abilities. His examples imply a preference to interpret the test as an function, or cognitive style, for subjects who are relatively field independent, with subjects' relative standing on other subtests indicating the degree to which they emphasize or deemphasize an analytical mode of processing spatial information. Wachtel appears to assume that the Spatial subtest scores of intelligent children should be as high as their Verbal Scale subtest scores. It is the present author's contention that interpretation of the Spatial or PO factor WISC-R subtests as a measure of field independence cannot be validly based on their standing relative to the Verbal Scale subtest scores. If a child has high scores on the Spatial subtests, he is probably both able and disposed to behave in a field independent manner; if not, he may be unable to disembed stimuli, indisposed to use an analytical style, or both. Further investigation of the hypothesis that a child's low Spatial subtest scores reflect field dependence should involve further testing with more direct measures of the FI construct. Both paper-and-pencil and kinesthetic FI measures could be applied to this investigation. The latter suggestion is partly based on the failure of the two types of FI measures to consistently form a unitary factor in the context of various intelligence, achievement, and perceptual tests. Since the Spatial subtests tend to correlate with both types of FI measures, the ambiguity regarding the relative importance of general intelligence and analytical style preference which surrounds FI measures is also present for interpretation of the WISC-R Spatial factor.

The field independence construct offered as an interpretation for the Spatial factor, and perhaps the entire PO factor, of the WISC-R has also been explored in some depth for its ability to explain the findings of cross-cultural comparisons of cognitive test results. The studies of this sort which pertain to Inuit and North

American Indian samples are reviewed in Chapter II.C. The focus of the present discussion of WISC-R interpretation shifts to the subtests within the FD factor.

Freedom from Distractibility Subtests

The third factor obtained from analysis of the WISC-R subtests has been has been interpreted and labelled in accordance with several theories, with no consensus in sight. Kaufman (1975) labelled the factor Freedom from Distractibility, as Cohen (1959) had labelled the similar factor of the WISC. Vance et al. (1978) interpreted the factor as a measure of short-term memory, labelling it Stimulus Trace. Goodenough and Karp (1961) found that all three of the corresponding WISC subtests loaded with tests involving memory and selective attention when factor analyzed with a larger battery of cognitive tests. Goodenough and Karp labelled their factor Memory-attention, although their discussion of the factor seems to emphasize the attention component. Bannatyne (1974) labelled the corresponding WISC factor Sequencing, suggesting that the common requirement underlying the subtests was the ability to process stimuli or responses in an imposed sequence. Kaufman (Note 1) later adopted the Sequencing label and interpretation, citing clinical observations that scores on the factor were often unreflective of the level of anxiety or attention displayed by children during testing. Kaufman's switch in factor labels also reflected adoption of Das et al.'s (1979) simultaneous-successive processing theory as a framework for interpretation. This theoretical stance led to the Naglieri et al. (1983) factor analysis of eight subtests which was discussed earlier in this section. The following discussion examines studies with Digit Span and Coding which might provide insight into the FD factor as a whole.

Digit Span was associated with Das's Successive factor in a study by Das, Kirby, and Jarman (1975), in which that WISC-R subtest loaded on a factor with a paired-associate learning task, a paper-folding task, and a test which required them to list as many potential uses for a target object as they could imagine. The paired-associate learning task involved pairs of abstract symbols. A similar task involving pairs of concrete objects did not load on this factor. It is also interesting to note that memory span for words did not load on the same factor as Digit Span. These results suggest that stimulus properties may [illegible] important to the definition of the factor underlying Digit Span, whether that factor is related to memory, mode of processing, or attention.

Dempster (1981) reviewed the literature pertaining to individual and developmental differences in memory span, examining both strategic and nonstrategic variables as explanatory factors. Although the implementation of strategies such as grouping items, chunking items into recognizable units, item rehearsal or item retrieval strategies was reported in response to many of the tasks employed in the studies, implementation of the strategies was not consistently related to individual or age differences in performance on digit span tasks. Much of the evidence which favored an explanatory role for strategic variables was collected from performances on supraspan tasks, i.e., long lists of items which are scored by counting the number of items recalled before the subject's first error. Results from experiments with supraspan tasks often failed to generalize to tasks like Digit Span, in which the subject must correctly reproduce the entire list of items to receive credit (Dempster, 1981, 1982).

Of the nonstrategic variables examined in Dempster's (1981) review, only ease of item identification was consistently related to individual and developmental differences on memory span. Ease of item identification was associated in most studies with speed of item identification, as measured by the minimal presentation time required for identification of a stimulus or by latency from stimulus presentation to oral naming of a stimulus. Some support was provided for an explanatory role for the ability to order items and mastery of the concept of order, but Dempster warned that this variable required more direct examination. Dempster also concluded that differences in the memory capacity required for mental transformations of stimuli, rather than developmental differences in the amount of memory capacity available, account for age-related differences in memory span for subjects older than 7 years. Susceptibility to interference was associated with individual and developmental tasks with some similarities to memory span tasks, but Dempster argued that this relationship was still speculative until supported by replication and direct experimentation with memory span tasks. A Similar conclusion was presented regarding the rate at which information stored in memory was searched. Dempster's overall conclusion was that the speed at which stimulus items are identified is the most important factor underlying individual differences in memory span tasks similar to the WISC-R Digit Span. He conceded that item identification speed, ability to encode order, and susceptibility to proactive inhibition might be jointly related to memory span or may share some common variance among themselves. He suggested that the relationship between speed of item identification and intelligence is directly related to the complexity of transformations on the item required before production of a response such as reproducing the list for the experimenter.

A series of experiments which examined the memory span of bilingual subjects provide some additional evidence for the relationship of speed of item identification to individual differences in span. Ellis and Hennelly (1980) found a negative correlation between the time required to read digit lists aloud, i.e., pronunciatiation speed, and digit span. Although subjects were generally more fluent in Welsh, they were able to pronounce digit names more quickly in English, a finding that Ellis and Hennelly attribute to phonemic properties of the Welsh names. The authors speculated that mental arithmetic calculation would be more difficult in Welsh than in English as a result. An alternative explanation is greater familiarity with English names for digits, reflecting their English education and almost exclusive use of English for calculation. This alternate hypothesis gains some support from the finding that subjects could read English and Welsh lists of names for the digits (as opposed to the digit symbols) at equal speed. Ellis and Hennelly also varied the language of digit presentation and recall in digit span tasks, finding that English presentation and response was associated with longer span recall. Lowest recall was achieved when the subject was presented the lists of digits in Welsh, their first language, and required to repeat the lists in English. The recall-inhibiting effect of translation was interpreted as an effect of memory space available for storage of digits being reduced by the working-memory demands of translation. The above results could also be interpreted as support for Dempster's (1981) claim that speed of item recognition is the critical variable for explaining digit span, as speed of pronunciation was related to span length. A similar relationship between Digit Span scores and speed of pronunciation of digit lists was reported for bilingual subjects tested in Chinese and English (Hoosain, 1982).

Studies which have examined the forward and backward components of the WISC-R Digit Span as separate tests have attempted to isolate the complexities of item transformations required by each component. Jensen and Figueroa (1975) argued that Digit Span Forward (DSF) is a measure of level I intelligence, involving rote memorization, while Digit Span Backward (DSB) is a measure of level II intelligence, requiring transformation of the series of digits before reproducing them for the examiner. The strongest support provided by these authors is the finding that DSB scores of a sample of California children had a significantly higher correlation with FSIQ than did their DSF scores. The other evidence provided by Jensen and Figueroa assumes adherence to Jensen's (1969, 1980) theory that black-white differences on intelligence test scores reflect inherited differences in level II intelligence. However, the hypothesis that DSB is more closely tied to general intelligence has received some additional support from Griffin and Hefferman's (1983) finding that, while WAIS DSB correlated significantly with VIQ, PIQ, and FSIQ, only VIQ was significantly correlated with DSF. The generalizability of Griffin and Hefferman's results to the WISC-R is limited by the fact that an adult psychiatric sample was employed.

The validity of interpretation of forward and backward span as separate tests has also been examined by searching for associations between DSF-DSB discrepancies and injury to specific areas of the brain. However, the relationship of DSF and DSB to localization of brain damage has primarily been examined with adults tested on the WAIS. Patients with right brain injuries and noninjured control subjects obtained DSF scores which were superior to those of patients with left brain injuries, while superior DSB scores were obtained by control subjects and those with left brain injuries (Weinberg, Diller, Gerstman, and Schulman, 1972). Weinberg et al. suggested that DSB performance difficulties experienced by patients with right brain injuries may be associated with difficulty in visual scanning and eye movement. The functional link between scanning and DSB may be explained by the hypothesis that subjects respond to the latter task by forming a mental image of the array of numbers as they are presented and then read the array from right to left. Costa (1975) provided some support for the hypothesis that visual scanning is related to DSB performance, as higher DSB scores were obtained by patients without neurological signs of visual field or visual attention deficits than by patients with such signs. However, DSF scores were not differentiated across groups and both left and right brain injuries were associated with DSB scores which were lower than those of control subjects. Costa concluded that DSB may reflect only generalized intellectual impairment. McFie (1975) suggested that the type of errors commited on Digit Span are indicative of the type of brain injury that might be present. General impairment may be reflected by perseveration, and McFie notes that a backward span more than three digits less than the forward span should raise the suspicion of general impairment of mental functioning. In summary, the research related to brain injury and DSF-DSB discrepancies provides some support that DSB is more closely associated with general intelligence, but also indicates that backward span may be affected by visual scanning deficits.

Although some authors have advocated separate interpretation of DSF and DSB (Gardner, 1981; Weinberg et al., 1972), others have cautioned that differences between the two Digit Span components are not reliable. Kaufman (1979a) calculated the average DSF-DSB discrepancy for the WISC-R standardization sample. On the basis of the distribution of discrepancies, he argued that the forward span should exceed the backward span by at least five or six digits before a deficiency in digit reversal is inferred. Inferences of deficiency in rote recall should require a backward span two digits longer than the child's forward span. Griffen and Hefferman (1983) note that the clinical utility of separate DSF and DSB scores is limited by the narrow range of such scores.

Digit Span's importance to the definition of the FD factor in Kaufman's model has led some authors (Evans & Hamm, 1979; Kaufman, 1982) to recommend that the subtest be routinely administered and not considered a supplementary test. However, the interpretation of that third factor is not clearly defined by the above research on Digit Span's relationship to variables extraneous to the WISC-R. A review of memory span research indicated that speed of item recognition was closely related to individual and developmental differences in length of span recalled. Dempster hypothesized that memory span's association with general intelligence would be a function of the complexity of mental transformations of the items required. That hypothesis received some support from findings that DSB, which requires the subject to reorder the items, is more closely related to general intelligence measures and from findings that generalized, or diffuse, brain injury is associated with low DSB scores.

A series of experimental manipulations of the task demands of the WISC Coding subtest and similar tests provide some insight into factors that Coding may share with Digit Span. Johnson and Lyle (1972a) hypothesized that poor coders were confusing reversible symbols included among the WISC digit-symbol pairs, i.e., pairs of symbols which are mirror-images of each other may be coded interchangeably by children with low Coding scores. However, the exclusion of such items from coding tasks benefitted both "good" and "poor" coders. This benefit reflected higher speed of production rather than a decrease in errors. Poor coders had more difficulty than good coders in matching the digit--symbol pairs from memory after one trial of Coding (Johnson & Lyle, 1972b). Although training subjects to increase recall by labelling the symbols led to increased recall for both good and poor coders, and posttest recall scores for the two groups did not significantly differ, the training did not narrow the Coding gap between good and poor coders. Writing speed, as measured by the number of "X"s placed in a paper grid within a fixed time limit, was found to be strongly related to Coding scores and to account for between-group variance on a figure copying task (Lyle & Johnson, 1973). Monitary reinforcement on a fixed-ratio schedule had no effect on either writing speed or Coding scores (Lyle & Johnson, 1973). Good coders were superior to poor coders in a paired associate learning task, as measured by the number of trials required for recall to the criterion level (Johnson & Lyle, 1973). The authors hypothesized that quick memorization of the digit symbol pairs would speed coding production, while frequent referral to the coding key would inhibit production. Teaching the digit-symbol pairs to criterion prior to testing plus removal of the key from the test paper resulted in relatively lower scores for both good and poor coders, contrary to expectations. The subjects appeared to waste more time trying to recall correct symbols for a given digit than they would have spent referring to the key. Johnson & Lyle (1973) concluded that difficulty in learning paired associations, which would normally be accomplished during the coding task, rather than before, would increase the frequency of glances at the key, thus slowing performance.

Lyle and Johnson (1974) entered the writing speed and paired-associate learning tasks from their earlier experiments into regression analyses to predict Coding scores for a new sample. Since females in their sample had obtained higher scores than males on all three tests, regression analyses were conducted separately for each sex, as well as for the total sample. Verbal and nonverbal IQ scores were also entered as predictors in all three analyses, while sex was included in the total sample analysis. Paired-associate learning speed and writing speed were joint predictors of Coding scores for both sexes, and sex contributed additional predictive power for the total sample. The contributions of the IQ tests were not significant. All the variables were included in a factor analysis which resulted in interpretation of three factors. The first factor was identified by the two IQ measures, while the other factors included paired-associate learning with Coding, and writing speed with Coding, respectively. Lyle and Johnson concluded that Coding scores were a result of the latter two orthogonal factors and that the effects of the factors were equally weighted. The high correlation with paired-associate learning has been replicated for the WISC-R Coding subtest with a sample of learning disabled children (Dean, 1983).

A clear definition of Kaufman's FD factor is not immediately apparent from the above review of studies which have attempted to identify the factors underlying tasks similar to Digit Span and Coding by experimentally altering the tasks demands. Some similarities in the results of the two sets of studies offer possible hypotheses. Although manipulation of task strategy has had demonstrated effects upon scores on both subtests, such manipulations did not diminish individual and developmental differences which were apparent for the original tests. Since writing is not allowed in Digit Span or Arithmetic, and since the latter subtest requires direct manipulation of objects on only the early counting items, writing speed (or even motor speed) can probably be dismissed as a candidate for the FD factor. Both Digit Span and Coding appear to involve short term memory. They require the subject to hold various items in memory while producing responses related to other items. Increasing the complexity of required transformations on the items (e.g., repeating digit strings backward or discriminating between mirror-image coding symbols), has a depressing effect on scores on both tasks.

The common factor may be the speed at which subjects learn associations (such as digit-symbol pairs, digit-serial position pairs, linguistic-correspondent pairs, etc.), and identify stimuli as members of a class or pair. In other words, borrowing from Demster's review, the FD factor may reflect efficiency in perceiving stimuli and transforming stimuli representations, so that the minimal amount of memory capacity is diverted from holding those item representations in short term memory. Performance on the Arithmetic subtest would conceivably be facillitated by speed of calculation (beyond that required to complete the item within the time limit) as slow mental calculations at one stage of item completion might conceivably increase the probability of forgetting other information relevant to that problem.

The above interpretation is highly conjectural, as interpretation of mental test factors has tended to be. Disagreements on the interpretation of both the PO and FD factors illustrates the inconclusiveness of interpretations based on armchair task analysis or examination of studies of each test within the factor. The difficulty with this approach is especially critical for the FD factor, which includes subtests such as Information and Coding inconsistently across age groups. Experimental manipulation of task demands of individual subtests probably helps to identify the factor by providing a more detailed examination of the nomological network in which those subtests are located. Correlations of factor scores with extraneous variables serves a similar purpose. However, ascription of meaning to the factor would be aided by the direct manipulation of task demands across all subtests within a factor, followed by direct examination of the effects of such manipulations upon factor patterns and, if the factor pattern remains intact, upon factor scores. For the present, the WISC-R appears to be comprised of three factors which vary in the amount of agreement voiced on their meaning. In Embretson's (1983) terms, the nomothetic span of the set of WISC-R subtests is fairly stable, whereas their construct representations are not clear. The following section describes methods for calculating scores for the three factors and then examines the relative validity of these scores, IQs, and Bannatyne categories for prediction and explanation of school-related problems. Methodological weaknesses of analysis of individual subtest profiles are also discussed, pointing to the importance of the factor models for construct validity of the WISC-R.

Derivation and Validity of WISC-R Factor Scores

Derivation of Scores for Kaufman's Factors

Three factor models for interpretation of the WISC-R have been discussed and evaluated through a review of factor analytic, simple correlational, and experimental studies with samples from both narrowly- and widely-defined samples. All three models have received qualified support. The Verbal and Performance Scales as two factors and Kaufman's VC, PO, and FD factors were replicated with various age groups within the standardization sample and some clinical and MR groups, although not without exception or occurrences of complex variables. The Spatial, Conceptualization, and Sequencing categories, which Bannatyne had derived from analysis of the WISC, received support in that the Sequencing category is identical to Kaufman's FD factor, while the subtests from the Conceptualization and Spatial categories loaded on the VC and PO factors, respectively, with the greatest consistency. The organization of the

Verbal and Performance Scales and Bannatyne's categories has been defined previously in this report. The method for calculating VC, PO, and FD factor scores is described below, followed by an examination of the criterion-related validity and clinical utility of scores calculated from each of the three models.

Several methods have been proposed for calculating WISC-R factor scores which would have means of 100 and standard deviations of 15. Sobotka and Black (1978) recommended summing the scaled scores of the subtests in each factor. The sums for the four subtests in each of the VC and PO factors (assuming that Mazes is not administered) would then be prorated according to tables available in the WISC-R manual (Wechsler, 1974) for that purpose. The prorated sums of scaled scores would then be converted to IQs, using Wechsler's tables for calculating VIQs and PIQs from these sums. The FD factor subtests would be summed and this sum prorated and converted to an IQ, using the corresponding tables for the WISC-R FSIQ. If Mazes were administered, the score on that subtest would substitute that of Coding, and the PIQ would then be calculated in the normal way and interpreted as a measure of Perceptual Organization.

Tellegen and Brigg's (1967) described a procedure for norming factor scores on the WISC which appears to be more straightforward, from a psychometric perspective, than utilization of the approximation methods of Sobotka and

Black (1978). Deviation scores on each factor were defined as

$$DQ = (15/Sc)(Xc-\bar{X}) + 100 \qquad (II.2)$$

where Xc is the sum of scaled scores, or composite score, for that factor, Sc was the standard deviation of the composite score, and $\bar{X}$ was the average composite score for the standardization sample. Sc was calculated from a formula which accounts for the variance of each subtest and for the intercorrelations among the subtests within a factor. Tellegen and Briggs provided tables of WISC composite score means and standard deviations for insertion into Equation II.2.

Gutkin (1978, 1982) has provided a linear regression equation for direct calculation of WISC-R deviations quotients (DQ) for the three factors (i.e., VCDQ, PODQ, and FDDQ), based on a mean of 100 and standard deviation of 15 for each factor. The regression equation was derived from the standardization sample data (Wechsler, 1974) and is based on the same properties of means and variances of composite scores as Tellegen and Briggs (1967) applied to derivation of WISC DQs. Use of the DQs, rather than IQs, altered the interpretations of WISC-R profiles for a sample of children referred for assessment (Gutkin, 1978) and for children in the standardization sample (Gutkin, 1982). ALthough the correlations between the IQs and their corresponding DQs were above .95, 15% of the standardization sample obtained either significantly large VIQ-PIQ discrepancies but nonsignificant VCDQ-PODQ discrepancies, or vice versa. Arithmetic scores were significantly discrepant from the average scaled score of the VCDQ subtests for 25% of that sample; Coding was discrepant from the PODQ subtest average for 28% of the children (Gutkin, 1982). Adherence to the three-factor model would therefore be expected to result in clinical judgements about the Verbal and Performance Scale tests that would differ from judgements derived from calculation and comparison of VIQs and PIQs. These differences would extend beyond the interpretation of an extra factor, as reflected in the shifts in Verbal--Performance discrepancies.

The reader may note that the above methods for calculating factor DQs from scaled scores involved an equal weighting of the subtests which defined that factor, rather than a direct application of the factor loadings as regression weights. The exclusion of variables with nonsalient loadings on a factor is supported by Horn's (1969) argument that the reliability of a factor is reduced by the contribution of variance from variables which are only randomly associated with the factor. In other words, variance from nonsalient variables is error variance. Horn cites evidence, from a thesis completed under his supervision, that factor scores based exclusively on salient variables are more stable in cross-validation than factor scores which directly apply the factor structure matrix. Gutkin (1979c) has calculated the reliability coefficients and standard errors of measurement (SE_m) for the factor DQs. The median SE_m values for the VCDQ (4.79) and PODQ (4.72) are comparable to those of the VIQ (3.57) and PIQ (4.65), which were derived from the same data and based on the same sample variance. The median SE_m for the FD factor is 5.62. Given that these factor scores are thought to be more factorially pure (Gutkin, 1982), one might have expected that the reliability of the factor scores would be higher than those of the VIQ and PIQ. The lower reliability, and consequently larger SE_m, for the FD factor is not surprising in light of the relative instability of that factor's definition and the large contribution of test-specific and random error variance to the subtests which load on that factor (Silverstein, 1976).

Kaufman (1979a, 1979b, 1982) has advocated a plan for clinical interpretation of the WISC-R which utilizes both the two- and three-factor approaches and allows analysis of subtest profiles under specified conditions. He argues that the Verbal and Performance IQs should be interpreted as the Verbal Comprehension and Perceptual Organization factors, respectively. The third factor should not be interpreted unless the following conditions are met: the scaled scores for Arithmetic, Digit Span, and Coding are all significantly different from the mean scores of their respective scales; those three subtests are sufficiently similar to be considered measures of a single trait. If a subtest score is significantly higher or lower (±3.0 points) than the mean scaled score of subtests in the same factor, Kaufman recommends labelling that subtest as a strength or weakness, respectively. Educational remediation may be based on such profile analyses. The general principle of Kaufman's plan is to operate from the hypothesis that verbal and nonverbal intelligence account for the subtest scores and that specificity of subtests or intermediate factors should only be considered after rejection of that hypothesis.

Kaufman's procedural guideline regarding the third factor reflects his acknowledgement that the FD subtests have tended to be complex in many factor analytic studies of the WISC-R, i.e., they have had loadings on the first or second factor as well as on the third (Kaufman, 1980). His procedure seems to imply that when the FD subtests are not clearly and consistently discrepant from the other two factors that the third factor underlying those subtests either has no influence upon the subtest scores or that its influence is not worth worrying about. In the present author's opinion, this logic reveals profound confusion regarding the definition and nature of a construct. Kaufman seems to imply that unidimensionality can be inferred from flat profiles. This argument, if taken to its logical conclusion, would state that a child's memory (or sequencing ability, anxiety, etc.) is not measured by the third factor unless the child's rank order on that ability is different than his rank order on verbal and perceptual abilities. The experimental research on Digit Span and Coding, which was reviewed earlier in this chapter, disconfirmed a unidimensional interpretation of those subtests. Equality of an individual's scores across the three subtests does not make the scores unidimensional or discount the influence of a specific factor(s) upon the composite score. Including the FD subtests in the calculation of the VC and PO factors, through interpretation of the IQs, will add variance attributable to the third factor to scores on the first two.

It is the present author's contention that clarity in interpretation of the WISC-R would be better served by calculating factor scores in the manner formulated by Gutkin (1978, 1982). The resulting VC and PO factors would be more homogeneous. If the subtests in the FD factor are widely scattered, analysis of the subtests, in the context of additional testing results, behavioral observations and interviews with the child, would be aimed at forming hypotheses to explain that scatter. The clinician may then decide that the third factor score is not a valid measure of memory or attention for that child without makng the erroneous assumption that those subtests are therefore valid measures of the first two factors.

The validity or clinical utility of the Verbal-Performance model, Kaufman's three factors, and Bannatyne's categories are examined below. Where Kaufman's factors were studied, factor scores were calculated as average scaled scores or in the manner suggested by Gutkin (1978).

Validity of the Interpretive Models

The following review of studies addressing applications of factor and subtest profile analysis is far from exhaustive. It is intended to highlight some of the threats to validity of the various models and consequent suggestions for further research. Applications of the Verbal-Performance model, Kaufman's three-factor model, and Bannatyne's categories are considered in turn. Methodological concerns which severely limit the validity of analysis of subtest profiles are then raised, stressing the importance of determining the validity of factor scores for psychological assessment.

Verbal-Performance Model

Studies of the relationship of WISC-R IQs to the Stanford-Binet Intelligence Test suggest that the VIQ and FSIQ are more closely related to the Binet test than is the PIQ (Raskin, Bloom, Klee, & Reese, 1978; Wechsler, 1974). Raskin et al. explain these results as a function of the Binet's dependence on verbal expression. The Raskin et al. sample consisted of children referred for diagnosis of mental retardation or a learning disability. Wechsler's sample was a small subset of the WISC-R standardization sample, tested at age 6 years, 9 years, 12 years, or 16 years. The correlations between the Binet and the VIQ ranged ,64 to .77 across the age groups; for the Binet and PIQ, .51 to .74; for the Binet and FSIQ, .63 to .82. These correlations are moderate, and, although they indicate that the Binet and WISC-R share some variance,the two tests do not appear to measure identical constructs.

Studies of the relationship of the Verbal and Performance IQs to school achievement tend to favor the VIQ as a predictor for most academic measures. The VIQ and FSIQ were highly correlated with all subtests of the Peabody Individual Achievement Test (PIAT) for an adolescent sample, while the PIQ was significantly correlated only with the Reading Comprehension and General Information subtests and the total PIAT score (Applebaum & Tuma, 1982). The PIQ of these children was correlated with Word Identification and Passage Comprehension on the Woodcock Reading Mastery Test, but regression analyses indicated that the PIQ did not contribute predictive power beyond that of the VIQ. Wikoff's (1979) results with a clinical sample partially contradict those of Applebaum and Tuma, as Wikoff found that all IQs were significantly correlated with all PIAT subtests. However, VIQ was still the best single predictor of all PIAT subtests, and PIQ did not contribute significant additional predictive power.

The samples in studies reported by Wikoff (1979) and by Applebaum and Tuma (1982) were of approximately average intelligence. The following studies used samples with lower mean FSIQs. Whereas the VIQ and Verbal subscales correlated with the Reading, Spelling, and Arithmetic subtests of the Wide-Range Achievement Test (WRAT) for a sample of children referred for institutional placement, their PIQs were uncorrelated with WRAT-Spelling and few of the Performance subtests correlated with either WRAT-Reading or WRAT-Arithmetic (Covin & Lubimiv, 1976). WRAT-Arithmetic, rather than Reading, was correlated with the PIQ in Hartlage and Boone's (1977) sample of public school children. Discrepancies between their tables and their citations and discussion in text obscure interpretation of Hartlage and Boone's findings for the VIQ and FSIQ relationships with the WRAT. All WRAT subtests were correlated with all WISC-R IQs for a second-grade sample tested by Hartlage and Steele (1977). The first- and second-grade school-assigned marks in reading, writing, arithmetic and social studies were correlated with VIQ. The PIQ was also significantly correlated to school-assigned marks in writing for both grades, and to arithmetic and reading at the first and second grades, respectively. PIQ was not correlated with social studies marks for either grade. with social studies and arithmetic marks. Since no information is available regarding the reliability of the grading system used in the schools, correlations with school performance must be cautiously interpreted.

The relationship of VIQs to academic achievement are consistently significant. The PIQ does not have relationships to academic variables which are as strong or as consistent as those of the VIQ. In some studies, PIQ has correlated with reading scores, while in others there has been a stronger relationship with Arithmetic. These mixed results for the PIQ may reflect differences in the definition of the samples, the integrity of the PIQ as a factor score, or restriction in the ranges of scores, as many of the samples in the above studies had mean FSIQs below 90.

A significant discrepancy between Verbal and Performance IQs has been suggested as a diagnostic indicator of a specific learning disability. In her review of the literature pertaining to LD diagnosis with the WISC, Lutey (1977) reported that PIQ exceeded VIQ for 77 of 81 samples of reading disabled children. Smith (1978) demonstrated the stability of this discrepancy over a seven-month period for a sample of children in classes for the learning disabled. Zingale and Smith (1978) demonstrated that the discrepancy was independent of socioeconomic status (SES), as measured by parental occupation and education. Low SES was associated with lower test scores, but the ranking of subtest or IQ scores was not affected by SES. Zingale and Smith conceded a point that is apparent from a survey of studies pertaining to the diagnosis of learning disabilities, i.e., the operational definition of disabilities varies across educational districts and governments, undermining any generalization of results with samples diagnosed as LD.

The validity of the VIQ-PIQ discrepancy as a diagnostic tool has been called into question by analysis of the frequency of significant discrepancies in the general population of school children. Kaufman (1976c) examined such frequencies in the standardization sample. He found that an IQ discrepancy of 12 points in either direction, which is significant at the .05 level of Type 1 error probability, was obtained by 30% of the children. Unlike Zingale and Smith's (1978) LD sample, the direction of the discrepancy was related to SES in Kaufman's sample. High SES children tended to have higher VIQs, while low SES children were more likely to have higher PIQs. Bloom and Raskin (1980) directly compared the frequencies of VIQ-PIQ discrepancies for the standardization sample to the frequencies for an LD sample. The mean discrepancy size for the LD sample was 10.6; for the standardization sample, 9.7. The frequencies of children with discrepancies at successiive intervals are closely paralleled across the two groups. Kaufman has suggested that only VIQ-PIQ discrepancies which are large enough to be rare in the normal population should be used for forming diagnostic hypotheses, such as the presence of a learning disability. However, Kaufman stated that a significant discrepancy may be used for identifying relative strengths and weaknesses and planning remedial education.

Kaufman's Factors

The relationships of factor scores to academic achievement was examined for a mixed clinical sample of students by Grossman and Johnson (1982). The sample had a mean FSIQ of 90.61 and age ranged from 6 years to 16 years. The VC and FD factors predicted scores on all WRAT subtests, with the FD factor score as the best predictor of each of

WRAT-Reading, Arithmetic, and Spelling. The PO factor was not significantly correlated with any of the WRAT scores. Whereas most scores for the WRAT and WISC-R ranged from 80 to 90, the PO factor and PIQ were 98 and 95, respectively. Grossman and Johnson noted that the ranking of factor scores, PO>VC>FD, was similar to the Bannatyne pattern associated with learning disabilities, i.e., Sp > Con > Seq.

The relationship of the factor scores to the achievement of educable mentally retarded children was tested by Cummins and Das (1980) after replication of the factors through factor analysis. Their results are somewhat different than those of Grossman and Johnson's sample, as VC scores were not correlated with WRAT scores. FD scores were strongly correlated with WRAT-Arithmetic. WRAT-Reading and Spelling were highly correlated with a battery of tests, including a Digit Span task, thought by the authors to measure Das et al.'s Successive Processing factor. The validity of these results is threatened by two characteristics of the study.

1. Only the ten mandatory WISC-R subtests were administered, so that Digit Span was not included in the FD factor.

The consequent weight of the Arithmetic subtest in the factor score may have inflated its relationship with WRAT-Arithmetic, while its relationship with WRAT-Reading and Spelling may have been deflated by the absence of Digit Span.

2. The sample FSIQs ranged from 55-80.
Kaufman (Note 1) has warned that IQ scores below 70 are not sufficiently reliable to justify discriminations between subjects in that range.
Correlations of scores in such extreme ranges are therefore of limited utility. Nonsignificant correlations may be affected by the restriction of scores and their low reliability.

Groff and Linden (1982) tested young (age 8 years to 11 years) and older (13 years to 16 years) samples of MR children on the WISC-R. They also tested a nonretarded sample which was similar in chronological age to the young MR sample; in mental maturity, to the older MR sample. While the mean VC, PO, and FD scores of the young MR children formed a flat profile, the PO scores of the nonretarded and older MR samples were higher than their VC or FD scores. The authors seem to imply a trend toward specialization with increasing mental maturity, but such a conclusion would require longitudinal comparisons of retarded and nonretarded children.

Bannatyne Categories

Based on a review of 22 studies of WISC patterns of disabled readers, Rugel (1974a) concluded that their Bannatyne scores were ranked with sufficient consistency to constitute a characteristic profile. That profile may be expressed as Sp > Con > Seq. A consistent profile did not emerge for the nondisabled samples.

Bannatyne added the Acquired Knowledge (AK) category in 1974, but as previously noted in this report, the validity of that category for the WISC-R does not have factor analytic support. Given that the category shares one subtest with each of the Con and Seq categories, comparisons of AK with either of those categories would have relatively low reliability. In replicating the Sp > Con > Seq profile for a sample of children diagnosed as LD, Smith, Coleman, Dockecki, and Davis (1977) found that AK scores were lower than Con scores, but similar to Seq scores. Lutey (1977) distinguished three LD patterns of WISC-R scores in a review of 30 studies. Eleven of these studies reported the Sp > Con > AK > Seq pattern, while ten studies reported Con as the highest score. The remaining nine studies reported that the SP score was ranked highest, while Con or AK were ranked lowest. Ascription of significance to such findings should probably await comparison to the frequency of such patterns which would be expected by chance alone.

Rugel (1974b) noted a tendency for disabled readers to have consistently lower scores on Information as well as on the Sequencing subtests. A pattern of low scores on these four subtests has been labelled the ACID profile (for Arithmetic, Coding, Information, and Digit Span). Ackerman, Dyckman, and Peters (1976) reported observation of the ACID pattern in the WISC scores of a group of LD children, without reference to the children's specific disabilities. They speculated that the relationship of Information to the

Sequencing tests may reflect difficulty in long term memory storage and/or retrieval. Lutey (1977) has generalized this profile to the WISC-R scores of disabled readers. She speculated that both Information and Arithmetic were measures of long term memory, while Digit Span and Coding measured short term memory, implying that reading disabilities reflected memory deficits. Although the experimental research on Digit Span and Coding, which was reviewed earlier in this chapter, identified associations for those tests with such short term learning components as speed of item recognition and speed of learning paired associations, respectively, Lutey's inferences regarding long term memory are conjectural at this stage. Although the ACID pattern has been reported for several LD samples, its diagnostic value is limited, as reflected in Ackerman et al.'s (1976) observation that Sequencing is rarely observed as a relative strength for any group of children.

The diagnostic utility of Bannatyne's categories was discredited by Clarizio and Bernard's (1981) failure to clearly distinguish groups of LD, emotionally handicapped (EH), educable mentally retarded (EMR), otherwise handicapped (OH), or nonhandicapped (NH) children. With the exception of the EMH, all group means reflected the pattern Sp > Con > Seq. However, only 35% of LD children would have been correctly classified with this criteria, while 32% of the NI children would have been misclassified as LD. Clarizio and Bernard concluded that Bannatyne's WISC-R categories were not effective for differential diagnosis, although they might be useful for planning remedial education.

Although patterns of Bannatyne scores do not appear to be useful for differentiating diagnostic groups at any given level of general intelligence, samples of intellectually gifted, average-IQ, and MR children can be differentiated on the basis of such patterns. Mueller, Matheson, & Short (in press) calculated mean Bannatyne scores for 36 samples from published studies on the WISC-R. A cluster analsis was conducted on the 36 profiles of means, resulting in the formation of three clusters, which corresponded to the gifted, average-IQ, and below-average samples. The pattern for the gifted group was Con > AK > Sp > Seq; for the average samples, Sp > Con > AK > Seq; for the below-average samples, Sp > Seq > Con > AK. The patterns for the gifted and below-average samples are almost mirror images of each other. The pattern for the average cluster was less consistent within the cluster than the patterns for the extreme groups. The four-cluster solution had split the average cluster into one cluster of nonhandicapped average-IQ children and one cluster of LD children plus nonhandicapped average-IQ children. These two average-IQ clusters did not differ in regards to the means of their category scores or the shape of their profiles. Mueller et al. concluded that Bannatyne categories were not effective for distinguishing diagnostic groups within restricted IQ ranges. A subsequent meta-analysis of 114 samples failed to find unique profiles associated with an extended set of diagnostic categories (including emotional disturbances), sex, or ascribed ethnic-group membership (Mueller, Note 2). Mueller's extended study replicated the differences between IQ groups reported by Mueller, Matheson, and Short (in press). The finding that profiles for gifted and below-average clusters show opposite trends may be a function of the relationship between FSIQ and the various category scores, i.e., scores having higher correlations with FSIQ would be expected to differentiate groups which have been differentiated on the basis of FSIQ. Subtests in the Con and AK categories had the highest such correlations for the WISC-R standardization sample (Wechsler, 1974). Scores on these subtests would therefore be expected to be the most extreme scores for extreme groups. Validation of this explanation is planned for the near future.

Profiles of Individual Subtests

Several guides are available to the clinician wishing to impute meaning to the differences among their clients' subtest scores. Sattler (1982) provides possible interpretations for various pairwise scaled score comparisons. Banas and Wills (1978) categorized subtests into components which each include one Verbal Scale test and one Performance Scale test. For example, Picture Completion and Information are labelled as tests of memory for isolated data; Similarities and Object Assembly, as tests of inferential thinking. The makeup of these components appears to have been based on the clinical experience of Banas and Wills, who propose teaching prescriptions on the basis of comparisons across and within components.

Lutey (1977) proposed the calculation of a series of supplementary scores, some of which are based on theoretical or intuitive grounds, rather than empirical validation. Lutey has renamed Bannatyne's Spatial category as Field Independence and the Sequencing category as Freedom from Anxiety. However, the result of the addition of the Information score to the Sequencing category is proposed as a measure of memory. Substituting Mazes for Block Design turns Lutey's Field Independence measure into her Freedom from Uncertainty measure. The validity of inferring such radical shifts in the meaning of factors underlying sets of composite scores which are almost dependent, i.e., sharing two out of three equally-weighted subtests, is extremely dubious.

The methods cited above for comparison of individual subtests or composites with shared subtests have not been empirically well validated. References are often made to theory or clinical experience of the authors. The reliability of such comparisons is also a cause for concern. Based on the premise that the reliabilty of a difference score will be positively associated with the reliability of the two tests, while negatively associated with their intercorrelation, Mueller, Mancini, and Short (Note 3) examined the reliability and efficiency of comparisons among subtests, among Verbal and Performance IQs, among Kaufman's factors, and among Bannatyne's categories. These authors applied Kelly's (1923) formula for calculating the proportion of observed test differences which exceed chance to the WISC-R reliabilities and intercorrelations for the standardization sample (Wechsler, 1974). They evaluated the results for each possible comparison against the criterion that at least 25% of observed differences should be reliable, i.e., exceed measurement error. All possible subtest comparisons failed to meet this diagnostic efficiency criteria on at least one age level. Eleven such comparisons were inappropriate at all age levels. However, comparisons between the Verbal and Performance Scales and among the Kaufman factors were appropriate, or efficient, at all age levels. All possible Bannatyne comparisons were efficient, with the exception of comparisons between Acquired Knowledge and either Conceptualization or Sequencing. Since AK shares a subtest with each of Con and Seq, lower reliabilities would be expected for such comparisons. Comparisons among scores on Kaufman's three factors were the most diagnostically efficient.

The diagnostic efficiency criteria employed by Mueller, Mancini, and Short (Note 3) assumed that the detection of reliable differences was important for identifying the strengths and weaknesses of the children assessed. The assignment of children into special diagnostic categories, such as mentally retarded or learning disabled, raises the concern that children not be erroneously assigned. Kaufman (1976a, 1976b) has suggested that diagnosis of a disability requires subtest scatter which is not only reliable, but occurrs rarely in the normal population of children. Silverstein (1981) notes that, given the correlation between the VIQ and PIQ, the size of discrepancy required to be in the extreme 5% of the population is approximately twice the discrepancy required to exceed the 95% confidence interval for the null hypothesis of a difference of 0. To aid the determination of the rarity of a given size of subtest scatter, Kaufman (1976a) provided normative tables on the frequencies of various sizes of the range of subtest scores and the number of subtests deviating significantly from the mean scaled score for the appropriate scale. Kaufman (1976a, 1976b) warned that ability profiles for normal children exhibit a great deal of scatter. The diagnostic utility of such scatter indices is undermined by Gutkin's (1979b) finding that children labelled emotionally disturbed, LD, brain-damaged, and educable mentally retarded could not be distinguished on the basis of subtest score range, whether measured across ten subtests or within the Verbal and Performance Scales, or the VIQ-PIQ discrepancy. All groups had significantly larger mean scatter indices than the standardization sample. Although this result suggests a role for scatter indices in detecting academic difficulty, the inability of scatter indices to distinguish among specific handicaps severely limits their utility for purposes of psychoeducational assessment.

The rank order of subtests in a child's score profile has been examined as a tool for differential diagnosis of cognitive and academic disabilities, with disappointing results. Hale (1979) classified a group of 8 year old children as underachievers in mathematics, in reading, in both subject areas, or adequate achievers, on the basis of their WISC-R IQs and WRAT scores. These groups did not differ on WISC-R FSIQ. A discriminant function was derived, and the application of that function resulted in significant variability on WISC-R composites across groups. However, the percentage of correct classification to groups was only 75.24, while the correct classification obtainable by simply assigning all children to the largest group was 72.8%. Only 29.2% of arithmetic underachievers were correctly classified, while no subjects with reading difficulties were correctly classified. Hale's findings support Berk's (1983) criticism that subtest profiles for diagnostic groups are not necessarily reflected by children within such groups.

In an attempt to integrate the literature on subtest profiles, cluster and profile analyses were conducted on the subtest means of 29 samples with below-average, average, or above-average IQs. (Mueller, Dash, Matheson, and Short, in press). As with the meta-analyses of Bannatyne profiles reported earlier (Mueller, Note 2; Mueller, Matheson, and Short, in press), distinctive profiles were found for the three IQ levels, but distinctive profiles were not found within any given IQ level. As with the Bannatyne meta-analyses, the profiles for the above- and below-average clusters reflect opposite trends, with the subtests most highly correlated with FSIQ ranked highest for the above--average samples; lowest for the below-average samples. Subtests with low correlations with FSIQ tended to have means close to the standardization sample mean of approximately 10. Where distinctive profiles do exist, such as for the gifted and MR samples in the meta-analyses described above, these profiles may simply reflect the varying communalities among subtests.

The low validity of WISC-R subtest or Bannatyne profiles for differential diagnosis has led several authors to strongly censure its use. Kaufman (1979a) branded differential diagnosis "based primarily on WISC-R subtest patterns a veritable impossibility" (p. 206). Vance, Singer, Kitson, and Brenner (1983) stated that, in the light of evidence against the validity of differential diagnosis, continuance of the procedure could be construed as malpractice. Hirshoren and Kavale (1976) argued that more research on profile analysis was required, but that its practice should cease.

If research on profile analysis continues, a number of weaknesses in the past and current research need to be addressed. The principal weakness is the reliability of the diagnostic criterion used in several studies to define diagnostic groups. Presence in classes for the LD, EMR, etc. is the operational definition of a learning disability or mental retardation. Criteria for diagnosis as disabled or retarded vary across states and provinces, making comparisons across studies invalid. While some studies (eg. Hale, 1979) distinguish among types of reading disabilities, others define learning disabilities or brain disorders within single overriding categories. While clinical neuropsychology texts, such as McFie's (1975), form hypotheses with the aid of the qualitative nature of errors on specific tests, Lutey's (1977) diagnostic guidebook advocated inferences on response hesitancy on the basis of the sum of scores on Picture Completion, Object Completion, and Mazes, which comprise three of the five PO subtests. In summary, much of the WISC-R profile analysis research has involved heterogenous, vaguely-defined samples. Other tests or observational measures have not been incorporated into the analyses, although Wechsler (1974) recommended that WISC-R interpretation be conducted in the context of such external measures. Bannatyne (1968) recommended a series of psycholinguistic and perceptual measures as part of a battery which would include the WISC category scores, but those measures were not included in attempts to validate the category scores as diagnostic indicators. Much of the research which examines the validity of profile analysis may be so far removed from the ideal practices of rigorous clinicians as to preclude the possibility of valid profile analysis. However, the research reviewed in this report strongly discredits diagnosis on the basis of profiles of WISC-R subtests alone, and presently offers no evidence to support diagnosis on the basis of WISC-R subtests plus some specific set of additional psychoeducational tests.

The two-, three,- and four-factor models of the WISC-R have had mixed success as predictors of school achievement. The VIQ and Kaufman's VC are consistently powerful predictors of academic achievement, whether measured by the WRAT or school marks. Kaufman's FD factor (or Bannatyne's Spatial category) has been associated with reading and arithmetic achievement. PIQ and the PO factor are less consistent predictors of achievement and tend not to add to the predictive power of the VIQ or VC factor, respectively.

The stability of factor analytic solutions for the WISC-R across methods and samples has led Hirshoren and Kavales (1976) to conclude that scores on three factors are sufficient to describe a child's performance on the test. Others (Kaufman, 1975, 1979a; Silverstein, 1977, 1980, 1982) have voiced support for both two- and three-factor models for interpretation. Although Bannatyne's original three categories received factor analytic support in that each is a subset of one of Kaufman's factors, the Acquired Knowledge factor is not empirically supported. The AK factor is not diagnostically efficient and its calculation and interpretation is not recommended.

Although differential diagnosis of learning disabilities and emotional disturbances has been discouraged, comparisons among factor scores are thought by some authors to have a potential role in guiding remedial education planning (Clarizio & Bernard, 1981; Kaufman, 1979a). These authors' discussions do not clarify the manner in which the label "weakness", as opposed to disability, will affect the nature and amount of remedial education required or offered. At present, performance on the WISC-R appears to be largely affected by three factors, although two factors can explain much of the variance. The nomological network surrounding the factors requires further clarification, and their utility beyond simple prediction requires more rigorous evidence than that available for diagnostic differentiation from subtest profiles. Pending the results of multivariate experimental research on factor scores, the role of the WISC-R in assessment should perhaps be as an initial step in a cognitive testing sequence, based on the Decision-making Model of assessment (Swanson & Watson, 1982) which was described in Chapter II.A. A significantly low score on a factor would lead to further testing with more simple measures of the components which had been found to contribute to variance on the factor (and therefore to the factor's operational definition) for children of the same age, educational, and linguistic background as the child.

In Chapter II.A, several definitions of test bias were offered. The test-bias definition of central importance to this report was identified as the failure of the test to measure the same construct across populations. The factor models for the WISC-R represent sets of hypothetical constructs to explain subtest performance. For example, Kaufman has identified three separate constructs within the twelve subtests and recommends interpretation. of the factors rather than the individual subtests in most cases (Kaufman, 1979a). The status of the factors as constructs may be threatened by lack of agreement about their nature, i.e., whether the FD factor represents memory, attention, anxiety, etc. Although there is still considerable debate about which cognitive processes or capacities are represented by the factors, they have been demonstrated to be more reliable and accurate predictors of academic performance than have separate subtests. The Verbal-Performance and Kaufman models have been replicated by a wide variety of methods with several samples. Generalization of the interpretation of WISC-R factor models to other normative populations would require a demonstration that the particular factor model explained the subtest intercorrelations obtained for those populations. If the factor models described above were replicated, such a generalization would also require experimental and correlational research with the factor scores in the setting of that new population. The next section of this report examines some of the literature pertaining to the bias of the WISC-R in several North American populations.

Bias of the WISC-R

This review of the construct validity of the WISC-R concludes with an examination of a small sample of the literature regarding the extent to which the test is biased for clinical or educational applications to members of various North American subgroups. Its reliability and factorial validity are emphasized. Most of the bias research deals with black and Chicano children in the United States.

Separate factor analyses for the black and white children in Wechsler's (1974) U.S. standardization sample resulted in closely-matched solutions for the two groups (Gutkin & Reynolds, 1981). The two-factor solution resembled the Verbal and Performance scales, except for Coding's low loadings on both factors. The three-factor patterns for the two groups were similar and replicated Kaufman's factors, although the third factor accounted for less variance for the black sample. Coefficients of congruence for corresponding factors were all above .98, leading Gutkin and Reynolds to recommend similar interpretive methods for black and white children.

Analysis of several other black samples has failed to find an interpretable or stable third factor. Sandoval (1982) factor analyzed the subtest correlations from the white, black, and Chicano standardization samples for

Mercer's (1979) SOMPA, a battery which includes the WISC-R. The third factor accounted for very little of the white sample's variance and did not emerge for the other samples. Sandoval rotated the two-factor solutions for the groups to achieve maximum similarity across groups. The resulting solutions were very similar, leading Sandoval to pronounce the WISC-R as unbiased for black and Chicano Americans. Silverstein (1973) had derived only two factors from the WISC data of white, black and Chicano children who had been referred for psychological services. Given the relationship of the FD factor to reading and arithmetic, and similar failures to find the third factor among clinic-referred children, as reviewed earlier in this chapter, the absence of this factor in Silverstein's results may reflect a restriction of range on those subtests. Since Sandoval's sample was a norming sample, based on California school children, his failure to find a third factor is a more important nonreplication of the factor analytic results for Wechsler's (1974) standardization sample.

Two-factor solutions were reported for the Chicano samples in the studies reviewed above. Gutkin and Reynolds (1980) reported similar findings for samples of white and Chicano children who had been referred for psychological services. They rotated a three-factor solution for both groups, but decided that the third factor was uninterpretable, although consistent across groups. Although the patterns of loadings in Gutkin and Reynold's samples are very similar, the sizes of the loadings and communalities are consistently smaller for the Chicano sample. Communalities for the two samples differ by .2-.3 for all subtests. This finding suggests that the factors are less stable for the Chicano sample.

Dean (1980) extracted and rotated three factors for each of two samples, labelled Anglo and Mexican-American, which had been referred for psychological evaluation concerning learning difficulties. All corresponding factors had congruence coefficients above .8, suggesting similarity between the factor solutions. However, matches between some of the noncorresponding factors (e.g., Factor II of the Mexican-American sample and Factor III of the Anglo sample) had congruence coefficients as high as .65, suggesting similarity among factors which should be orthogonal. The legitimacy of the application and comparison of congruence coefficients is obscured by ignorance about the distributional properties of the statistic. This problem and its importance for studies of test bias is discussed in more detail in Chapter II.D.

Reschly (1978) applied principal components, principal factor, and maximum likelihood factor analysis to the WISC-R subtest correlations for four samples. The samples included children in Grades 1 to 9 in Arizona schools, labelled Anglo, black, Chicano, and Native American Papago. The results for the Anglo group replicated Kaufman's (1975) original analysis, with interpretable two- and three-factor solutions. Two factor solutions for the other groups replicated the Verbal and Performance Scale organization of the test, especially if Mazes were substituted for Coding. The three factor solutions of Reschly's Anglo and Chicano samples closely resembled Kaufman's, but the third factor was considered uninterpretable for the black and Native samples. Reschly (1978) concluded that calculation of Verbal and Performance IQs would be appropriate for all groups. Scores on the Kaufman factors were significantly related to Metropolitan Achievement Test (MAT) scores for these samples (Reschly & Reschly, 1979). VC scores and FSIQ were the best predictors of both MAT-Reading and MAT-Math, followed by FD scores. All correlations were lower for the Native sample, leading Reschly and Reschly to conclude that the WISC-R--achievement relationship had been confirmed for all samples, with the exception of the Native Papagos.

Sheik and Miller (1978) reported a replication of the Kaufman factor model for a sample of children from low--income families in the southeastern United States. The sample included both white and black children, with an average IQ which was significantly below the standardization mean of 100. Oakland and Feigenbaum (1979) attempted to examine test bias for the WISC-R across age, ethnic, income, and other variables, using the SOMPA standardization data. However, their claim that the test has construct validity for all groups across all classifications is based on their finding that VIQ, PIQ, FSIQ loaded together on a factor with

Bender Gestalt, when included in the factor analysis of a battery of educational, medical, and demographic measures. Since FSIQ is almost completely jointly-determined by VIQ and PIQ, this finding provides no evidence that the test or scales measure any specific construct.

Gutkin (1979a) calculated the reliability of three Bannatyne (1974) categories for a sample of Mexican-American children who had been diagnosed as LD. The split-half reliabilities for Conceptualization, Spatial, and Sequencing were .92, .89, and .86, respectively. Gutkin concluded that the categories were not sufficiently reliable to justify their use in educational placement decisions for Mexican--American children. His criterion of adequate reliability was a coefficient of .9. Since the FSIQ mean for Gutkin's sample was 76.6, the reliability coefficients may have been restricted due to a restriction of the variability of scores. Furthermore, Kaufman (Note 1) cautioned that WISC-R scores are less reliable as they approach the extremes of the distribution. Since the mean score for this sample is 1.6 standard deviations from the standardization mean of 100, the reliabilities would be expected to be low. Dean (1977) reported split-half reliabilities of .65 to .93 for individual subtest scores of a Mexican-American sample. These reliability coefficients were larger than those reported by Wechsler (1974), although not significantly so. The FSIQs for Dean's sample ranged from 86 to 108.

The concern regarding the construct validity of the WISC-R for American minority groups reflects an awareness of a pattern of lower mean scores for many of those groups. Kaufman and Doppelt (1976) divided the standardization sample for the WISC-R on the basis of several demographic variables. They found differences in FSIQ means across regions, with the western and northeastern states reporting the highest IQs; the southeastern states, the lowest IQs. Children of professionals obtained higher IQs than children of laborers. Urban children scored higher than rural children, although the mean difference was just over two points. Black children scored approximately 15 points below white children, on the average. Black urban children scored an average of 4.5 VIQ points and 2.6 PIQ points higher than rural black children. Vance and Engin (1978) reported higher IQs for males than females in their sample of rural, black children, while Vance and Gaynor (1976) found that urban, black females obtained higher VIQs than urban, black males. Although some inconsistencies regarding the interactions among such demographic variables are reported, and although the distributions for the various groups overlap, high--income, urban dwellers of European or Asian ancestry have obtained consistently higher average scores than low-income, black or Chicano rural dwellers.

The trends described above have provided ammunition for all sides in nature-nurture and test bias debates. Whereas Jensen (1980) has concluded that most individually-administered intelligence tests, including the WISC-R, are not biased for American-born speakers of English, Block and Dworkin (1976) have called for a moratorium on the testing of children from minority groups. Kaufman (Note 1) has tended to concur with Jensen's conclusion, with one qualification. Native North Americans, according to Kaufman, are the one American minority group for whom the statistical indicators of test bias have been consistently positive, i.e., for whom differential predictive and factorial validity have been demonstrated. He based this qualification on results such as Reschly's (1978) failure to find an interpretable FD factor in the WISC-R scores of Native Papagos in Arizona, and the low correlations between WISC-R factor scores and achievement for that sample (Reschly & Reschly, 1979). The following section of Chapter II reviews the literature pertaining to trends in test scores for various Native Canadian and American samples. Explanations which have been offered for these trends are examined, and methodological issues which confound such explanations are noted.

C. Psychoeducational Assessment of Native Children

As stated in the previous section of this chapter, statistical indices of test bias have not consistently implicated the WISC-R for clinical use with most American minority groups. Native Americans have been identified as an exceptional group, in that bias indicators have been consistently positive for Native samples. This section of Chapter II reviews the literature on the patterns of cognitive and perceptual test scores of Native samples, in search of common trends which might offer hypotheses on the operational nature of such biases. The WISC-R is examined first, followed by a discussion of the validity of other tests. Bilingualism and qualitative properties of some Native languages are identified as confounding variables in the assessment of Native children. The impact of certain health problems on test scores is explored, along with those of technological and social change to lifestyles in Canada's eastern arctic.

Discussion of research involving Native samples throughout Canada and the U.S. is not intended to imply generalizability of the findings across all groups of Native children. Where ecological variables have been carefully examined, these are highlighted. However, systematic examination of the effect of such variables is rare among published studies, in spite of the lip service paid to acknowledging diversity among Native communities. Confusion regarding the definition of Native culture and rapid change in the social and economic ecology of eastern arctic settlements further confound interpretation of the test scores of children from these communities.

The WISC-R: Score Patterns and Validity Indices

Studies which have directly examined the utility and validity of the WISC-R for Native children have been largely restricted to Navajo and Papago children in the southwestern U.S. Verbal Scale subtest scores have been consistently lower than Performance Scale subtest scores. The mean VIQ and FSIQ for one primary-grade Navajo sample were reported as 64.14 and 77.06, respectively, while their mean PIQ was 95.41 (Hynd, Quackenbush, Kramer, Conner, & Weed, 1979). The latter score was well within the average range, as defined by the U.S. standardization sample. Teeter, Moore, and Petersen (1982) compared Hynd et al.'s results to those for 6 year to 16 year old Navajo children who had been referred for psychological services. Those children who had been diagnosed as nonhandicapped (NH) obtained scores on the Spatial subtests which equalled or exceeded the normative average. Children diagnosed as educationally disadvantaged (ED), i.e., with low academic achievement attributed to environmental or experiential factors, also performed within the national average on the Block Design and Object Assembly subtests. Children in both of the above groups obtained higher PIQs than the children labelled LD, but all groups had extremely low VIQs.

The score patterns exhibited on the small sample of WISC-R studies have been obtained with Native samples for the other Wechsler tests. Lutey (1977) reviewed studies of fifteen Native samples who had been administered either the

WISC, WAIS, or Wechsler Preshool and Primary Scale of Intelligence (WPPSI). Median VIQ-PIQ discrepancies for the three tests were 28.3, 25.2, and 13.9 for the three tests. King (1967) reported that Native children attending a regular classroom in the Yukon achieved above-average scores on the WISC Performance subtests; below-average scores on the Verbal subtests. Taylor and Skanes (1975) noted that the WPPSI Performance scores of kindergarten and first-grade children in Labrador were higher than their Verbal score~~s~~, although even the former scores were well below the standardization mean of 100. Similar results were found for the WISC scores of Native children in Montana, who scored below a white comparison sample on all IQs and all subtests except Block Design and Object Assembly (Peck, 1973). St. John and Krichev (1976) tested 100 children, aged from 6 years to 15 years, on the WISC and an additional 33 subjects, 16 to 20 years, on the WAIS. Their sample was comprised of Ojibwa and Cree children from northern Ontario. The VIQs for the sample ranged from 69.7, for the 6-7 year old children, to 91.1, for the 18-20 year old subjects, while the PIQs for these groups ranged from 99.58 to 103.4. The mean VIQs of the age groups were positively and significantly associated with increasing age.

The Bannatyne patterns of Native children who had been diagnosed as LD have not followed the Sp > Con > Seq pattern thought to reflect learning disabilities. McShane and Plas (1982) found a Sp > Seq > CON = AK pattern of WISC-R and

WISC subtests to be consistent for a sample of Ojibwa and Sioux children referred for assessment of learning disabilities, for assessment of hearing difficulties, and for giftedness screening. When their sample was split into groups, labelled as acculturated and traditional on the basis of such indices as Native language fluency, attendance at Native religious ceremonies, or the size of their Verbal-Performance discrepancy, the differences among category scores were only significant for the traditional group. McShane and Plas labelled that profile an "Indian pattern". However, the use of VIQ-PIQ discrepancies to determine level of acculturation would have spuriously produced differences between Spatial and Conceptualization scores for the traditional group and suppressed them for the acculturated group. The acculturation hypothesis is further discredited by examination of the actual mean scores obtained for the two groups. While the mean scaled score obtained on WISC-R Spatial subtests by the traditional group was 11.6, the corresponding mean for the acculturated group was 8.86. The failure to replicate the "Indian Profile" for the latter group reflects their relatively lower scores on Spatial subtests, rather than the higher Conceptualization scores which might have been predicted for an acculturated group. With the exception of the children referred for giftedness screening, it is not evident from McShane and Plas's report that children in the acculturated sample were functioning well in either a Native or nonNative culture, since the group is defined by the absence of a VIQ-PIQ discrepancy, and characterized by the absence of parental participation in Native ceremonies or fluency in the Native language.

Connelly (1983) reported the Sp > Seq > Con pattern in the mean category scores of an Alaskan Tlingit sample of children, aged 11 years to 16 years, who had been referred for assessment of learning problems or giftedness. Although the category rankings for 61% of the children in this sample corresponded to that profile, only 8% of the children exhibited significant differences in all of the pairwise comparisons which defined the Indian pattern. A younger sample (aged 6 years to 10 years) from the same school districts did not replicate the Indian pattern in the profile of category means. Although the mean Spatial score was significantly higher than the other category means, the Conceptualization mean score was nonsignificantly higher than that of Sequencing. The Indian pattern was reflected in the category rankings of 33% of the younger sample, but only 2% exhibited significant differences which were consistent with the pattern in all pairwise comparisons of categories. Connelly noted the age differences in the profile of means but placed more credence in the finding that the Indian pattern was replicated by more children in each sample than he claims would be expected by chance.

Examination of Connelly's means, standard deviations, and proportions for the various WISC-R scores and patterns indicate that his two samples differ in ways which should have prompted further analysis, research and discussion on his part. The proportion of older children whose category score rankings replicated the Indian pattern was significantly larger than the corresponding proportion for the younger children. The variances of VIQs, PIQs, and FSIQs for the younger sample were significantly larger than the corresponding variances for the older sample. The mean Conceptualization score of the younger sample was significantly higher than that of the older sample. The source of these cross-sectional differences is not discernible from Connelly's data, but an examination of such potential sources as school curriculum changes or recent additions to local educational resources (such as educational television) would be warranted.

The pervasiveness and uniqueness of the Sp > Seq > Con pattern for Native children is dubious. Connelly (1983) admitted that his results may have been affected by the fact that all but seven of his subjects had been referred for assessment as a result of school problems. McShane and Plas (1982) had also used a referral sample. Since children who are learning disabled or mentally retarded tend to score highest on the Spatial category (Bannatyne, 1974; Mueller, Matheson, & Short, in press), a large proportion of children in referred samples obtain their highest score on Spatial. Spatial had the highest rank for 100% of the MR samples and 91% of the low-average-IQ samples in Mueller et al.'s meta-analysis of 36 samples. The distinctiveness of the Indian pattern from Bannatyne's LD pattern is the reversal of the ranks of Sequencing and Conceptualization. Connelly appears to have calculated the chance expectancies for replication of the Indian pattern (8%) on the basis of all possible permutations of Sp, Seq, Con, and AK and compared his obtained proportions of occurrences of Sp > Seq > Con >AK and Sp > Seq > AK >Con to that figure. Since his sample was a referral sample, the calculation of chance probabilities should have accounted for the fact that Spatial would be expected to have the highest ranking for most children. If the 91% figure from Mueller et al.'s study is used as a rough estimate of the probability that Spatial would rank highest, the probability of an occurence of the Indian pattern in a referral sample due to chance becomes .30. This figure is almost exactly equal to the proportion obtained for Connelly's younger sample. If Acquired Knowledge scores are not considered in profile interpretation, as suggested in Chapter II.B, the probability of occurrence of the Indian pattern in a referral sample becomes .5.

Zarske and Moore (1982a, 1982b) replicated the Sp > Seq > Con pattern with Navajo children in both LD and regular classroom settings. However, they warned that the use of English as a second language may have confounded the FSIQ and Verbal Scale scores. The appearance of the Sp> > Seq > Con pattern in the WISC-R profiles of bilingual nonNative groups (Cummins, 1982; Gutkin, 1979a) undermines McShane and Plas' (1982) interpretation of the profile as a reflection of Native culture. Pronouncement of a Native profile on the basis of the pattern of means for one sample of clinic-referred children was an extreme overgeneralization, with or without evidence for a language-familiarity interpretation.

The reliabilities of WISC-R subtest scores and IQs of a sample of Navajo children were computed by Mishra and Lord (1982). They found that the split-half reliability of the PIQ was .86, while those of the VIQ and FSIQ were .63 and .72, respectively. The latter two reliability figures are in contrast to the corresponding U.S. standardization sample's average coefficients of .94 and .96. Seyfort, Spreen, and Lahmer (1980) found that the rank ordering of item difficulties within some subtests, as calculated from the scores of British Columbian Native students, did not correspond to the order of their administration. Since the test administration procedures require discontinuance of most subtests after some criterion number of successive failures, changes in item difficulty may lower and restrict subtest scores, thereby limiting reliabilities and intercorrelations.

In concluding Chapter II.B, it was noted that the Kaufman factor solution did not emerge for a sample of Native Papago children (Reschly, 1978) and that factor scores based on the Kaufman solution had weaker associations with academic achievement than was observed for children grouped as black, Chicano, or Anglo. Factor patterns for a Navajo sample and a Papago sample were interpreted by Zarske, Moore, & Petersen (1981) as evidence for interpretation of the WISC-R as a measure of verbal and nonverbal intelligence. Zarske et al.'s procedures preclude detection of developmental trends, as they analyzed the scaled scores for all children, from ages 6 years to 15 years, together. Naglieri (1982) has also criticized Zarske et al.'s conclusion, arguing that, although IQ-like factors may emerge, the factors do not necessarily measure verbal and nonverbal intelligence. Since English was a second language for the children, the factors may simply measure fluency in English. Zarske et al. (1982) qualified their conclusion, calling the factors measures of verbal and nonverbal intellectual abilities. They also argued that replication of the factor model upon which a clinical application may be based is a necessary but nonsufficient test of the battery's construct validity.

The Kaufman-factor scores of Native Papagos were demonstrated to be poorer predictors of academic achievement, as measured by the MAT, than were the corresponding scores for samples of black, Chicano, and Anglo children (Reschly & Reschly, 1979; Reschly & Sabers, 1979). Mishra (1981) found that the VC scores of fourth- and fifth-grade Navajo children were significantly correlated with WRAT-Reading, while FD scores were correlated with

WRAT-Arithmetic. However, these correlations were only .26 and .25, respectively, leading Mishra to conclude that WISC-R factor scores have limited utility for educational assessment with Navajo children. The predictive validity studies conducted by Reschly, Mishra, and their respective colleagues must be interpreted with caution, as the possibility exists that the WRAT and/or MAT are not valid measures of school achievement for the samples they studied. This is the problem of a biased criterion, which was discussed in Chapter II.A.

Almost all of the published research into the validity of the WISC-R for Native North American populations has been conducted in Navajo and Papago communities in Arizona. This research has discredited the WISC-R as a psychoeducational assessment instrument for children in those communities and schools. Very little information is provided about the lifestyle of those communities, so that generalization to other Native populations is largely guesswork. Reschly and Sabers (1979) note that the Natives in their sample were rural dwellers, while Mishra (1981) informs the reader that his Navajo sample attended a reservation school. Zarske et al.'s (1981, 1982) sample spoke English as a second language and their schooling was exclusively in English. Geographic isolation, rural lifestyles and livelihoods, segregation, and bilingualism are among many explanations offered for the patterns observed in cognitive and perceptual test scores of Native children.

While the research on WISC-R patterns and validity has been largely restricted to a small geographical region, research has been conducted on other test batteries with Inuit children and other Arctic samples. Some of this research is reviewed below. Generalizations and hasty conclusions have been offered about culture and language development as explanatory variables, and some of these generalizations are noted and critiqued. Also included in this review is research with southern Native and nonNative samples that has attempted to systematically define and observe those cultural variables.

Other Cognitive and Perceptual Tests

Scoring Trends

The verbal-nonverbal test performance discrepancy which was evident in the Wechsler scores of Native samples in the above review has also been observed in the scores of Native samples on an assortment of cognitive, academic, and perceptual tests. However, the discrepancy does not apply to all Native groups. The studies reviewed below found differing score trends across samples. These studies also strongly suggest that the dichotomization of tests into those requiring extensive verbal expression and those requiring perceptual and motor ability is too simplistic to explain patterns of Native children on psychological tests.

Taylor and Skanes's (1975) study of Labrador coast children, which was cited earlier in respect to the WPSSI, also examined scores on the Peabody Picture Vocabulary Test (PPVT) and the Illinois Test of Psycholinguistic Abilities (ITPA). The verbal-nonverbal model received some support from their finding that the Inuit children scored relatively low on such subtests as Verbal Expression, Grammatic Closure, Auditory Association, and Sound Blending, while achieving their highest scores on Visual Sequential Memory, Visual Association, and Visual Closure. Kleinfeld (1970) tested Native secondary students in three Alaskan boarding schools on the Academic Promise Test. Their average performances on the Numerical, Verbal, and Language Usage subtests were at the 24, 26, and 35%ile levels, respectively, of the U.S. norms for their ages. However, their average score on the Abstract Reasoning subtest, which uses geometric figures as task materials, was at the 49 %ile of the national norms. Vernon (1966a, 1966b) tested Inuit children in the McKenzie River delta region of the Northwest Territories and Indian children in Alberta. He reported low scores in tests from his *v:ed* factor, such as Arithmetic, while scores on a series of Piagetian concept formation tasks were similar to English norms. Poorer scores were obtained for Piagetian conservation tasks and an inductive reasoning test which employed numerical and verbal task materials. Vocabulary scores were closer to national norms when multiple choice items were employed than when the children were required to generate definitions. Inuit children achieved higher scores than Indian children on a test called Information Learning, which involved answering oral questions about text material which had been read aloud by the examiner. Vernon attributed this group difference to the Inuit children's more prevalent use of English at home.

Based on the assumption that group differences on tests of intellectual potential reflect test bias, MacArthur identified Raven's Standard Progressive Matrices (SPM) and Colored Progressive Matrices (CPM), Safran Culture-Reduced Intelligence Test (SCRIT), and the Spatial and Nonlanguage subtests of the California Short-form Test of Mental Maturity (CTMM) as least biased among a battery of intelligence and educational aptitude tests (MacArthur, 1968; West & MacArthur, 1964). Lower scores were obtained on tests requiring more verbal production on the part of the examinee, such as the Logical subtest of the CTMM and the Otis Quick-Scoring Mental Ability Test.

Scoring trends on the Raven's tests have been inconsistent across Native samples. Unlike MacArthur's Alberta sample, children in Wiltshire and Gray's (1969) sample from northern Saskatchewan obtained Raven's scores which averaged more than one standard deviation below the British average scores for their age level. Taylor and Skane's (1976a) sample of Inuit children on the Labrador coast obtained CPM scores which exceeded those of the white children in the same region. Feldman and Bock (1970) analyzed the SPM scores of children and adults in an Alaskan village and found that their achievement varied with characteristics of the items. The SPM is comprised of five sets of 12 items each. The early items in each set of twelve are relatively easy, and are intended to allow the examinee to form a problem-solving strategy for that set. Successive items involve more complex stimuli and are increasingly difficult. Successive sets also increase in difficulty (Raven, 1960; Raven, Court, & Raven, 1977). Feldman and Bock found that the rank order of difficulty level for sets was altered for their Inuit sample. They suggested that higher scores were obtained on set B than set A because the former required analogical reasoning, while the latter was more related to esthetic form, which may vary across cultures. Higher scores on set D than set C were explained by noting that set C is comprised of permutation problems, which Feldman and Bock argued are rarely learned unless specifically taught. Feldman and Bock noted that the norms on sets B and D for their Inuit sample were equivalent to Raven's Scottish norms, although their SPM total scores were well below the Scottish averages.

The Goodenough Draw-A-Man (DAM) test has been administered to several Native samples, with a strong trend toward average scores which match or exceed those of U.S. national norms. Mean DAM IQs for 6 year to 10 year old children from Sioux, Hopi, Zuni, Navajo, and Papago villages in the U.S. ranged from 99.2 to 133.8 (Havighurst, Gunther, & Pratt, 1946). Wiltshire and Gray's (1969) northern Saskatchewan Cree sample obtained mean DAM IQs of 113.5 for boys and 100.4 for girls. The sex effect was significant. Carney and Trowbridge (1962) reported that DAM scores of children from a Fox reservation in Iowa exceeded the national averages across all age groups from 6 years to 13 years. Findings of Native superiority on the DAM are not without exception, however. Comparison of various Native and white samples in the region of Vancouver Island, the British Columbia coastline, and an unspecified urban centre revealed no significant group differences on the DAM (Gaddes, McKenzie, & Barnsley, 1968).

Bland (1970, 1975) reported that Alaskan Native children and Native children from the south-western U.S. obtained higher scores on a visual memory task than white children from the same communities. Memory scores increased with age, but the gap between groups did not vary with age (Kleinfeld, 1971b). The items on Bland's tests were adapted from the Bender Gestalt Test. Kleinfeld (1973b) suggested that these results indicated that Inuit children would be particularly proficient at such SI factors as Memory for Figural Units, Evaluation of Figural Implications, and other tests with figural content. Kleinfeld argued that the development of skills for spatial perception and problem--solving were essential for survival in the Arctic. Findings of such spatial skills are not universal among Native samples, as Lowry (1970) found that her Native sample in southern Iowa obtained below-average scores on tests of visual perception and auditory discrimination.

The relationship of age to test score patterns is inconsistent across samples and tests. Scores on the CTMM verbal subtests regressed upward, toward the U.S. national averages, with increasing age for Carney and Trowbridge's (1962) Iowa Native sample, while their nonverbal subtest averages regressed downward, toward the U.S. average, over that same age span. As reported earlier, DAM scores for that sample remained superior across the age span studied.

Some of the samples in the studies reviewed thus far have achieved average or above-average scores on some "nonverbal" tests while scoring much more poorly on others (Wiltshire & Gray, 1969; Feldman & Bock, 1970). Such inconsistency may reflect differences in complexity of the instructions for the various tests. Feldman and Bock (1970) reported low scores for Alaskan children on a task which required mental rotation of a two-dimensional representation of a three-dimensional figure. The correlation of this task with English vocabulary was much higher than expected, leading the researchers to suggest that difficulty with the English instructions may have limited performances. Preston's (1964) Alaskan adult sample found the Picture Completion and Picture Arrangement subtests of the Wechsler-Bellevue (Wechsler, 1939) to be baffling, while scoring within the U.S. average range on Block Design and Object Assembly. (The Wechsler-Bellevue is the precursor to the WAIS.) This discrepancy would not be expected from the strong associations evident among Picture Completion, Block

Design, and Object Assembly scores of children in most samples, as reported in Chapter II.B.

Validity Indices

Vernon's (1966a, 1966b) assessment of NWT Inuit and Albertan Indian children reflected the relative strength on nonverbal tests which was suggested by much of the above research. Separate factor analyses of his battery for Inuit and Indian children resulted in some differences in the structure reported for those two groups (Vernon, 1969). While the general factor for the Inuit sample was mainly comprised of spatial tests, nonverbal intelligence tests, and the total score on the Piagetian tests, *g* reflected verbal, personality, and social variables, such as SES, to a greater extent for the Indian sample. Vernon noted that the record of Indian students indicated a more regular history of school attendance than was evident for the Inuit children. However, the Indian sample had a wider scatter of English proficiency and greater diversity of linguistic background.

Vandenberg and Hakstian (1978) reanalysed Vernon's data from Canadian Native groups, as well as data which Vernon (1969) had reported for boys in the Scottish Hebrides and in Uganda. Vandenberg rotated the factor matrices for the Canadian and Ugandan samples to that of the Hebridean sample by oblique Procrustes (which is described in Chapter II.D of this report). He reported that the similarity of the latter three group matrices to the Hebrideans' matrix was satisfactory, although dissimilarities were reported among those three groups. Hakstian rotated all matrices to a common target, using the method of Meredith (1964) (also described in Chapter II.D). He reported that four factors were interpretable across groups. These were: Achievement, which include the academic tests which Vernon labels *v:ed*; Ideational Fluency, which included projective tests; Conservation, which was dominated by Piagetian tests; and Spatial-perceptual, which included a number of drawing tasks, Raven's Matrices, Kohs Blocks (from which the WISC-R B.D. is adapted), and Porteus Mazes (the source of the WISC-R Mazes subtest). One major discrepancy among group factor patterns was the fact that *g* was primarily a spatial factor for both Inuit and Indian samples; a verbal factor for Hebrideans and Ugandans. The amount of English spoken at home correlated with the *v:ed* factor for the Hebridean sample, while correlating with the Ideational Fluency factor for the other samples. Vandenberg and Hakstian suggest that the Achievement factor reflected a mixture of fluid and crystallized intelligence for the Hebrideans, while measuring crystallized intelligence for the other samples.

MacArthur has found that factor patterns for some Canadian Native samples differed from those of nonNative samples. MacArthur's Native samples were Inuit children from the McKenzie delta region in NWT and Indian children from Alberta. He derived the factor patterns for these two samples on a large battery of cognitive tests, and compared these to the factor patterns of white children from the McKenzie delta (MacArthur, 1969). The age range of MacArthur's samples was approximately 9 years to 12 years old, which encompasses the age range of Vernon's (1966a, 1966b) samples from the McKenzie delta region. Factors labelled *v:ed* and Reasoning from a Nonverbal Stimuli (RNS) were identified for all three samples. An additional factor, labelled Verbal Memory, was identified for the white sample. While Raven's Matrices (SPM) loaded on the RNS factor for both the Inuit and white samples, this test loaded on the *v:ed* factor for Indian children. This difference must be interpreted with caution as some of the *v:ed* tests for the latter group were substitutions for achievement tests used in the NWT samples, inserted to accomodate local testing requirements. MacArthur (1973) compared the factor patterns from his battery for older Inuit children from NWT and Nsenga children from eastern Africa. He replicated the RNS loading for Raven's SPM with the Inuit sample, but this test loaded on the *v:ed* factor for the Nsenga sample. MacArthur has interpreted such differences in the associations of SPM, which is considered by many researchers (Jensen, 1980; Vandenberg and Hakstian, 1978) to be a good measure of general intelligence, in the context of ecological theories regarding the development of field independence. This interpretation is briefly described later in this section of Chapter II.

Bowd administered a series of tests of spatial ability and motor dexterity, along with SPM, a vocabulary test, a questionnaire pertaining to demographic variables, and tests of specific mechanical knowledge to Indian children from northern Alberta, central Alberta, coastal British Columbia and to white children from an Albertan urban centre. Discrepancies were evident among the four factor patterns. For example, a separate dexterity factor emerged for the coastal Indian and white samples, but not for the Alberta Indian samples. The general factor for the central Albertan sample was more heavily dominated by spatial tests than was true for other samples.

The tendency for Raven's SPM to load apart from other *v:ed* tests for several Native samples (MacArthur, 1969, 1973; Vandenberg & Hakstian, 1978) raises the question of the utility of this test for predicting the school performance of Native children. The evidence for SPM's predictive validity is not consistent across Native samples. Rattan and MacArthur (1968) reported that SPM, along with SCRIT, CTMM-Spatial, and other culture-reduced tests in MacArthur's battery were as effective in predicting academic achievement four years hence for his Albertan and NWT Native samples as were most of the more traditional tests requiring more verbal activity from the child. In contrast, Bowd (1972) found that the SPM scores of his coastal B.C. and central Albertan samples did not correlate with grade level, unlike the relationships found for Calgary whites and northern Natives. Grade level and vocabulary scores were significantly correlated for the three Native samples, but not for the white sample. Bowd interpreted his results as evidence that English vocabulary was a better predictor of academic achievement than intelligence for Natives in some remote communities. As with other attempts to determine the predictive validity of a test, the reader must be aware that the criterion test (the academic achievement test in this case) may be biased for the population of interest.

The results of validation studies on tests such as the SPM lead to a number of conclusions concerning the validity of cognitive tests for psychoeducational assessment of Native children.

1. Such tests are not necessarily equally valid across Native populations for any given purpose, as their correlations with school achievement variables are not totally consistent.
2. Children in the Native communities sampled tend to obtain low scores on tests which require the generation of extended answers in English, although they may obtain scores at or above the national average on tasks with similar content but different response modes.
3. Notwithstanding the above point, the Verbal-Nonverbal distinction is a simplistic description or explanation of Native children's performances on cognitive tests. Group profiles on sets of nonverbal tasks have shown considerable scatter (e.g. Wiltshire & Gray, 1969).

Discrepancies among Inuit, Indian, and white samples' factor patterns for batteries of academic, linguistic, and/or spatial tests, plus Feldman and Bock's (1970) findings regarding the rank order of difficulty levels for SPM sets, suggest the operation of factors other than verbal intelligence or spatial ability upon the patterns of cognitive test scores of Native children.

The present author's overall conclusion from a review of studies of the Wechsler tests, Raven's tests, and the other psychoeducational batteries discussed above, is that those tests can not be assumed to measure the same constructs for Native children as they have been assumed to measure in the larger population of North American school children. Some of the research which has attempted to explain the scoring patterns of Native children is reviewed below.

Inferences on Cognitive Processing

The patchwork of test results for Native samples in the above review might be expected to foster caution in generating theoretical explanations for the performance of any sample of Native children, or individual Native child, on the tests commonly used for psychoeducational assessment. Such has not always been the case. The following studies have made inferences on the cognitive and linguistic processing deficits or strengths of Native children from comparison of test performances across groups. These studies vary extensively in the precision and care with which samples are collected and defined. The methodological weaknesses of these studies will be briefly discussed, leading to a discussion on alternate explanations for such group differences.

The low scores attained by Native children on standardized tests of verbal skills have led some investigators to conclude that Native children are less able to use language to understand or express concepts (Mickelson & Galloway, 1969, 1973). Downing, Ollila, and Oliver (1975) offer the insights that Native parents are "less conscious of the act of speaking" (p. 313) and provide less verbal stimulation to the child at home, without identifying the evidence for these claims.

Mickelson & Galloway (1969) suggested that the language deficit is cumulative over the school years, based on the comparison of scores on an English language test between preschool children in a language training program and older children who had not received the program. Those authors felt that the gains achieved through training were greater than the gains achieved over time by the older children. However, the the use of the older children as a control group was inappropriate for numerous reasons. Pretest scores were not available for those children, so that the researchers do not know the extent to which their English language skills improved over time. The older children may have entered school at a different level of readiness than children in the experimental group. Mickelson and Galloway's study included several other threats to internal validity, of both design and measurement natures, yet their results formed the basis of strong conclusions about the nature of English language difficulties experienced by Native children.

Knowles and Boersma (1971) suggested that Native children were handicapped in tasks requiring verbal mediation of behavior, even when the target behavior is nonverbal. These researchers administered a discrimination task to 60 Native children from reservation schools and 30 white children from urban schools serving high-income families. Although a correct response to a change in the target behavior, i.e., reversal shift rather than a non-reversal shift, was made by a larger proportion of the white children than was true of the Native sample, this difference was nonsignificant. The researchers' hypothesis that occurrence of a reversal shift would be related to English language development was confirmed only for the white sample. They note that the Native children's first language was Cree, which may have affected that association. Native students who were prompted to verbalize and explain their choices were more likely to make a reversal shift, but this treatment may have only forced attention to the task. Knowles and Boersma's suggestion of a Native handicap in tasks requiring verbal mediation is remarkable for several reasons. They appear to ignore the fact that most of the group differences obtained were nonsignificant. Their white and Native samples differ on so many variables (urban vs. rural dwelling, family income, first language, attendance at a preschool, etc.), that attribution of even significant results would have been guesswork.

Schubert and Cropley (1972) examined the issue of verbal mediation of nonverbal tasks somewhat more systematically than Knowles and Boersma (1971). Their samples included Native children from reserves in northern and southern Saskatchewan, as well as white children from rural and urban communities. The population of the northern reservation, which had been inaccessible by road until 1965, made their living primarily from trapping and fishing. The southern reservation was close to an urban centre and social assistance was the main source of income. The children from the northern reservation achieved lower scores on a discrimination task than southern Natives, although they were able to learn and perform the discriminations equally well. The former group, who spoke Cree predominately, were penalized for difficulties in explaining the discrimination rules they had applied. The southern Natives, who spoke English predominately, were able to both apply and verbalize the correct rule at rates which were very close to those of the white samples. Schubert and Cropley noted that northern Native children with low IQs were able to improve their WISC Block Design and Similarities scores, as measured by a test-teach-test procedure, to a significantly greater degree than was observed for white children with similar IQs.

Northern Natives also outperformed white children of similar IQ level on the verbal regulation task. Schubert and Cropley emphasized the differences between scores of northern and southern Natives in concluding that low Native IQs have environmental causes.

Schubert and Cropley (1972) appear to make an assumption that failure to describe the verbal mediation that one employs in a task implies that the mediation is inadequate. Although this assumption seems reasonable, differences in the status of English across the samples may invalidate it. The northern Natives were required to express the rule in their second language, which may not have been the language in which they mediated the task. Their inability to provide satisfactory definitions of the rules they employed may have reflected difficulties in translation of the rule from Cree into English. If one wishes to accept the possibility that the Northern Natives correctly applied the discrimination rule without verbal mediation of that behavior, it follows that children in the southern samples may have done likewise, verbally formulating the rule after the fact for the experimenter. Consideration of such alternate explanations for Schubert and Cropley's results reveal the tenuous status of inferences about the employment of cognitive strategies from global test achievement, even with the aid of self-report measures.

The relative unfamiliarity with English which was characteristic of children in Schubert and Cropley's (1972) northern sample may also have led to difficulties in understanding the instructions of the various tests. Under such circumstances, their improvement in Block Design and Similarities may simply reflect their new understanding of the requirements of the task, gained from observing the experimenter in the teaching segment of the assessment. Rather than an effect of "environment" or, more specifically, of isolation from urban industrial society, the low IQs of the northern Native groups may be artifacts of fluency in English and test-wiseness.

Improper selection of cognitive strategies has also been implicated as a source of Native children's difficulties on cognitive tests. Das and Krywaniuk (1972) administered a number of tests from their battery of measures of simultaneous and successive processing to children on a Cree reservation in Alberta. They noted that the children did extremely poorly on a series of memory tasks which involved lists of semantically similar words and lists of phonetically similar words. They concluded that these children were using simultaneous processing strategies to solve the memory span tasks, which would be more effectively solved by successive processing. They administered a training program to teach the children to apply successive strategies to the task. The children were divided into two groups, with one group receiving more intensive training. Significant improvement in memory span was observed for the group receiving more training. The factor analysis results of the posttest scores for the pooled groups were more in line with Das et al.'s (1979) theory than those of the pretest results.

Das (1980; Das & Krywaniuk, 1972) presented these results as evidence that the Native children's test and achievement scores could be improved by training them to select simultaneous or successive strategies appropriately. Their conclusions are not supported, however, as the study suffers from numerous misapplications of experimental design and factor analytic methods. For example, by pooling the scores of the two experimental groups for the pretest and posttest factor analyses, they have precluded the possibility of testing whether the change in factor structure across trials was due to the experimental treatment. However, the main purpose in discussing this study pertains to an alternative explanation for the low scores in memory span at pretest.

Das and Krywaniuk noted that the children had difficulty in final consonant discrimination, leading to errors in repeating the words from the memory span list. Soveran (n.d.) has described differences between the sound systems of Cree and English speech which would contribute to such difficulties. Some pairs of English consonant sounds, such as "s" and "sh", "p" and "b", "k" and "g", and "ch" and "j", are distinguishable only by the effect of voicing, i.e., the tongue and lip positions for "p" and "b" are the same, but the vocal chords are active only in the production of the "b" sound. The auditory distinctions produced by voicing are not relevant to the semantic meaning of spoken Cree, as intonation is not important to the semantic meaning of spoken English. Hence, English word pairs which differ only in regard to consonant pairs such as those listed above will be indistinguishable to Cree speakers. Soveran provides examples of this effect on the spoken English of children whose first language is Cree. Of the twenty-four word lists on one of Das and Krywaniuk's memory span tests, twelve contained word pairs which would be indistinguishable for speakers of Cree. Order reversals could simply reflect a failure to discriminate between such consonant pairs. Discrimination difficulties could also have been reflected in the production of words which were not in the original list. Das (1980) has stated that the children in that Alberta sample spoke English habitually, and that only the elderly people in the community were able to speak fluent Cree. However, auditory discrimination difficulties were noted by Das and Krywaniuk (1972) and it is possible that those difficulties could reflect a Cree influence upon their English speech. Burnaby (1982, p.21) makes reference to "Indian English", a distinctive manner of speaking English which reflect influences of the Native language spoken in the community. Changes in the children's memory scores may reflect shifts in attention to phonetic properties essential to English, rather than shifts from simultaneous to successive processing strategies. The theory that Native children do not employ, or employ deficiently, successive processing strategies is further undermined by Das, Manos, and Kanungo's (1975) finding that children from the same Alberta reservation obtained mean scores on a visual sequential memory task which were equivalent to those of Edmonton school children.

The ambiguity of the meaning of Das and Krywaniuk's (1972) results illustrates the tenuousness of inferences about a child's cognitive processing strategies or capacity from global test scores, particularly where the child lives in a multilingual environment. Stronger evidence of the potential effect of selective attention to inappropriate phonetic properties, in the manner suggested as an alternative explanation for Das and Krywaniuk's results, was provided by a series of experiments by Serpell (1968). Serpell has demonstrated that Zambian speakers of Bantu, who experience discrimination difficulties with "l" and "r" which are similar to the voicing discrimination difficulties experienced by Cree speakers, produce errors on spelling tasks, matching tasks, and multiple choice tasks which reflect failures to attend to such discriminations. Serpell's findings challenged the belief that Bantu children's reading skills were similar to those of retarded native speakers of English. Serpell's emphasis on selective attention toward certain phonemic distinctions provides an interpretation to counter Lowry's (1970) assertion that Indian children are less developed in auditory discriminations than white children.

The most systematic research undertaken to explain Native children's psychological test profiles has focused on their area of relative strength. Canadian Native samples have been included in studies of the relationship of field independence to various sociological variables. Berry (1966, 1974, 1976) administered a series of FI measures to children and adults in urban and rural settings in NWT, Australia, New Guinea, Scotland, and Sierra Leone. The societies were chosen to provide a range over dimensions of ecological exploitive patterns (subsistence hunting vs. agriculture), the complexity of societal structure (small autonomous groups vs. hierarchical structure with clear status differentiation), and acculturation toward European lifestyles. Traditional Inuit society, as represented by Pond Inlet, NWT, was structured in small nomadic bands which gathered together at various times of the year, sustaining themselves through hunting and fishing. Children were encouraged to explore and become independent. The Temne of Sierra Leone lived in agricultural societies with rigid hierarchies and status differentiation. Child-rearing tended to be very strict. Other societies included in the study fell between these extremes on the ecological and complexity variables. Acculturation was varied by comparing isolated rural settlements, such as Pond Inlet, to urban centres within the same society, such as Frobisher Bay, NWT. Analysis of the FI measures supported Berry's hypotheses and

Witkin's (1974) theory on the relationship of socialization practices to field independence. Field independence was associated with loosely structured societies and with nomadic hunting lifestyles. Thus, the Scottish and Inuit samples achieved the highest mean scores on all but one test. The effects for acculturation were generally nonsignificant across all tests.

It is interesting to note that, while Raven's SPM, the EFT, Koh's Blocks, Morrisby Shapes, and a discrimination task were all strongly correlated with years of formal education for Frobisher children, corresponding correlations were nonsignificant for the Pond Inlet sample. This finding may reflect the development of field independence through exposure to formal education in one society; through the fostering of exploration and hunting skills in the other. However, since a school had been established in Pond Inlet only a few years before Berry's data was collected, the lack of association between schooling and FI measures may simply be an artifact of restricted variance for years of schooling. Information on the use of alternative schooling facilities, such as boarding schools in other communities, prior to the establishment of the local school, would have clarified the differences between the two Inuit communities.

Berry (1976) examined his theories further by administering FI measures to children in Indian communities across Canada which varied on the ecological and acculturation dimensions, although not to the same degree as the international samples had varied. Field Independence scores again favored the more loosely structured, hunting-dependent societies, such as the James Bay Cree samples, over the more hierarchical, agricultural societies, such as the Pacific coast Tsimshian. Acculturation was not significantly related to FI scores.

The multivariate experimental approach to construct definition suggested by Messick (1972) is evident in MacArthur's (1967, 1969, 1973, 1975a, 1975b) joint use of factor analysis and cross-cultural comparisons to examine the construct of field independence. MacArthur tested Inuit children in the McKenzie delta region of NWT, Indian and white children in Alberta, and Nsenga children in eastern Africa. The communities sampled were chosen to provide variation in child-rearing practices and ecological exploitation. The lifestyles of MacArthur's Canadian Native samples were similar to those of Berry (1966, 1976); of his African samples, to those of Berry's Temne samples. He used Vernon's (1965) hierarchical factor theory of intelligence to interpret the factor analytic results, noting that FI measures loaded with other spatial, or *k:m* tests. The results of cross-cultural comparisons were very similar to those of Berry in that rural North American Native samples scored close to urban white samples on spatial tasks, although scoring lower on verbally-loaded tasks. The Nsenga sample obtained lower mean scores on the spatial factor as well as the verbal factor. The separate factor defined by the scales of Raven's Progressive Matrices and interpreted as Reasoning from Nonverbal Stimuli (RNS) was unique to the Inuit children, whose mean scores on Raven's scales were close to that of the white sample. These scales loaded with *v:ed* tests for other samples, suggesting that the lifestyle of the Inuit children in the sample may foster cognitive skills which are acquired through more formal schooling in other societies.

The cross-cultural studies of Berry and MacArthur are not free of threats to internal validity. Variables such as ecological exploitation and child-rearing practices tend to vary together, i.e., they are not independent of each other. This correlation is consistent with the theories of the researchers and those of Witkin, but the design of the above studies precludes the reader from determining if the variables could be separated. These variables may be accessible only as ex post facto variables, rendering such separation impossible or impractical. However, this research represents the most systematic attempt to date to identify the interactions among specific variables contributing to the development of cognitive skills in Native Canadian children. The authors' conclusions were offered in the context of their concept of intelligence as adaptation to one's environment and to changes in that environment (Berry, 1976). The Inuit people's highly developed skills for articulation, or analysis of figures within an ambiguous background, are a response to the tasks required for survival in a setting where people must range over a large, relatively featureless, terrain. Travel in small bands and encouragement of individual exploration and independence are seen to foster the necessary skills (Berry, 1966, 1976; MacArthur, 1969, 1975a). Carpenter's (1955) observations on the mechanical and map-making skills of the Inuit of Southampton Island reflect a concept of perceptual flexibility which is very similar to definitions of field independence offered by Berry and MacArthur. Carpenter referred to a tendency to avoid conceptual separation of space and time, i.e., to conceive of space as a direction, rather than an enclosed area. Similar observations have been made about the orientation toward time and space reflected in Inuit languages, such as Inuktitut (Brody, 1977; Denny, 1973; Gagne, 1968).

Many of the studies pertaining to Native children's score patterns which have been cited thus far have offered assessments on the levels of verbal concept development (Mickelson & Galloway, 1973), verbal mediation ability (Knowles & Boersma, 1971; Schubert & Cropley, 1972) and literacy awareness (Downing et al., 1975) of Native children on the basis of their performance in English. Of the studies available to the present author, only the Alaskan research of Feldman and Bock (1970) compared the results of a test as administered in both English and the appropriate Native language. The following section raises some issues pertaining to bilingualism among Native people and the relationship of bilingualism to intelligence test scores. The purpose of this discussion is to give the reader a glimpse at the complexity of the network of variables affecting the relationship.

Bilingualism and Psychological Assessment

Earlier in this report, it was stated that the verbal--nonverbal dichotomy of tests was too simplistic to explain the patterns of test scores obtained by Native children. It was also stated that inferences about verbal cognitive processes and capacities of bilingual children could not be validly assessed in their second language alone. The present section is concerned with the ways in which scores on psychoeducational tests may be biased as a function of test administration in the child's second language. The relationship of bilingualism to scores on psychological tests is briefly described, leading to the conclusion that the relationship interacts with many social and linguistic factors. The languages spoken by Canada's Natives are surveyed, focusing on the arctic regions and the languages of the Inuit. Finally, the potential for characteristics of the languages themselves to interfere in the test achievement of bilingual children will be explored.

Bilingualism and Intelligence Scores

Reviews of early studies on bilingualism and intelligence indicated a trend for bilingual children to have below-average scores on verbal tests (Darcy, 1953, 1963). These children were particularly penalized on timed tests and group-administered verbal tests. When unilingual and bilingual groups were not matched for SES, nonverbal scores also tended to be lower for bilingual children (Darcy, 1963). Most of the studies in Darcy's reviews were conducted with Spanish-American children in Puerto Rico, New York, and the southwestern U.S. The nature of the samples and the tests used was such that the effect of bilingualism on intelligence could not be determined. The phenomenon being observed was the scoring trends on tests in the majority language for people who spoke a minority language as their first language.

The trends described above have been reflected in the IQ and subtest profiles of the Wechsler test scores of children speaking English as a second language. Corwin (1965) matched Mexican-American children, whose first language was Spanish, on CTMM-Total IQ with children whose first language was English. The WISC Verbal Scale and Full Scale scores of the former group were lower than those of the latter group. The children speaking English as a second language also had lower subtest scores for Similarities and Vocabulary.

Cummins (1982) calculated the median WISC-R subtest scores of Canadian immigrant children in English as a Second Language (ESL) courses in Toronto. The median Performance Scale subtests were all ranked higher than the median Verbal Scale subtests. Cummin's graph of the medians suggests that if factor scores were calculated from Kaufman's model, these would be ranked as PO > FD > VC; or Spa > Seq > Con, using Bannatyne's categories. This Bannatyne configuration was evident for the Mexican-American sample of Gutkin (1979), the Navajo sample of Zarske and Moore (1982b), Connelly's (1983) older Tlingit sample, and for McShane and Plas' (1982) sample, which was predominantly Ojibwa and Sioux. Unless the variance of subtest scores for the ESL sample was very large, the graph suggests that the VC score could be significantly lower than the PO and FD scores, which would almost certainly not be significantly different from each other. What McShane and Plas (1982) referred to as an "Indian pattern" might be called an ESL pattern with as much validity, and greater parsimony. Consideration of the low diagnostic utility of WISC-R profile analysis, as discussed in Chapter II.B, does not foster high hopes for the clinical utility of the pattern's detection, regardless of the interpretation. However, this remains an empirical question.

The meaningulness of the above pattern of medians is limited by two possible artifacts of a characteristic of Cummin's (1982) sample.

1. Medians were reported because not all of the subtests were administered to all children. Subtest means, based on children with complete data, may be quite different from the medians presented, depending on the shape of the subtest distributions and the score characteristics of children omitted by such a process.

2. The fact that Digit Span was administered to only 104 of the 242 children in Cummin's sample suggests that the tests may have been administered in response to referrals for assessment. If the sample is collected from school referrals, the referral process may have favored the selection of children with low Verbal subtest scores.

The rank order of the FD subtests in the WISC-R subtest profiles of Cummin's (1982) sample of immigrant children may simply reflect greater familiarity with English names and symbols for numbers than with English words and sentence structure. In Chapter II.B, the discussion of factors affecting scores on Digit Span included references to research which compared the digit spans of bilingual adults in Welsh and English (Ellis & Hennelly, 1980) and in Chinese and English (Hoosain, 1982). Ellis and Hennelly had suggested that the shorter pronunciation time for English digit names resulted in conservation of working memory capacity, allowing longer spans in English than in Welsh. However, the samples in both studies had had extensive schooling in English. It was noted by the respective authors that their subjects had not had much exposure to arithmetic outside the school setting, and were consequently more comfortable dealing with numerical material in English, rather than their first language. Similar observations were made for a Hebridean sample of children by Smith and Lawley (1948). Gaelic was the language used by these islanders in all transactions outside of school. However, during administration of intelligence tests in Gaelic, they often switched to English to solve arithmetic problems. The three bilingual samples discussed above were similar in that, in each case, the children's first language was that of the majority of people in the community, although much of their schooling was in English (perhaps less so for the Hong Kong sample). This characteristic is shared by Inuit children in many settlements in Canada. Superior scores in the FD or Sequencing factor of the WISC-R, relative to the VC factor, may reflect greater familiarity with the stimuli. The confounded inferences regarding Native sequencing ability which Das and Krywaniuk (1972) proposed on the basis of phonetically ambiguous word lists highlight the importance of the nature of the test stimuli for the assessment of bilingual children.

The trend toward lower verbal scores for children learning English as a second language is neither constant across social conditions nor necessarily permanent across the life span. Cross-sectional analysis of the Ammon Picture Vocabulary Test scores of immigrant children in Toronto ESL classes suggested that approximately five to seven years was required for these children to approach grade norms in the English application of academic skills (Cummins, 1982). Cummins (1978, 1979) has proposed a threshold hypothesis, which states that a minimum proficiency is required in the first language to avoid negative effects from the introduction of a second. He suggested that children speaking a minority language are often required to replace it with the majority language, although the former is not adequately developed.

The importance of the majority or minority status of the first language was reinforced by Toukomaa and Skutnabb-Kangas' (1977) studies of Finnish-speaking children in the Swedish school system. Attendance at a Finnish nursery school and instruction in Finnish in the early years of elementary school gave these children an advantage in linguistic development over Finnish-speaking children who began their schooling in Swedish. These authors classified programs for teaching a second language as: immersion programs where the second language is introduced as enrichment; submersion programs, which are aimed at the assimilation of speakers of a minority language; and emancipatory programs, which use the minority language as the sole medium of instruction in the early grades and both the minority and majority languages as mediums of instruction in the later grades. They noted that immersion programming tends to work well for introducing a second language to students whose first language is the majority language. The first language is not threatened in that situation, as the children have opportunities to develop their first language outside the classroom. Unlike speakers of minority languages, they are not required to replace their first language. Toukamaa and Skutnabb-Kangas recommended that the introduction of a second language as a medium of instruction be delayed until the age of 10 or 12 years, to prevent the discontinuity of language development they observed among minority group children.

Research into the relationship of bilingualism to cognitive development must account for the minority and majority status of the children and the languages involved as demonstrated by Cummin's Canadian research and Toukamaa and Skutnabb-Kangas' studies with children in Sweden. The relationship may be much more complex than even that body of research suggests. By mapping out the possible interactions among various political and social factors, linguistic policies of government, relationships between the languages involved, and prevalence of bilingualism among speakers of each of the languages, Gaarder (1977) has identified 54 possible outcomes for the development of the minority and majority languages, i.e., 54 types of bilingualism. This complexity confounds attempts to derive valid conclusions from the literature pertaining to bilingualism and the cognitive development of Native children. The configuration of relationships among the languages spoken by Canadian Natives is also complex. A brief description of this configuration is provided below.

Native Languages in the Canadian Arctic

There are currently 11 language families among the Native languages in Canada. The Algonquian languages are the ancestral language of the largest number of Native people.

Including Cree, Blackfoot, and Micmac, among several others, these languages are spoken in all provinces, although British Columbia has few communities where an Algonquian language is spoken. At the other extreme, some linguistic families include only one Native language, such as Kutenai or Haida, which are spoken by less than 2000 people each. Most of these isolated linguistic groups are spoken in communities on the Pacific coast or the southern interior of British Columbia (Burnaby, 1982; Statistics Canada, 1971).

The NWT population is largely comprised of Dene, Inuit, Metis, and EuroCanadians. The languages spoken include 8 Dene languages, 19 dialects of Inuktitut, English, and French (Watters, 1980). The Dene are comprised of the Indian bands in NWT, which live primarily in the forested areas in the southwest portions of the territory. The Dene languages belong to the Athapascan linguistic group, which also includes languages spoken by Natives in the northern sections of the four western provinces. Watters noted that the Slavey language alone, which is among the Dene languages, has 14 dialects. Inuktitut, which is included in the Eskimo-Aleut family, is the ancestral language of the Inuit people of Canada. Not all dialects of Inuktitut are interchangeable. Gagne (1968) has claimed that the largest gap is between the dialects of the McKenzie delta and the eastern arctic. This linguistic mosaic is further complicated by the development of mixtures such as Deltanese. This combination of English, Inuktitut, and the

Dene language, Loucheux, is spoken in the McKenzie delta region (Scollon, 1979).

Accurate information on the use of Native languages is more difficult to obtain than population figures for each ancestral language group. The federal government's Department of Indian Affairs and Northern Development (DIAND) reported in 1980 that 31% of Indian children spoke only English upon school entry; 23%, a Native language only; 35%, a Native language plus English; 2%, a Native language plus French; .8%, French only; .25%, a Native language plus English and French; 8%, unknown. Since these figures are taken from the DIAND registry of Indians, they would not include Metis, Inuit, nonstatus Indians, or several Dene bands. Burnaby (1980) analyzed national census data for 1971, focusing on those people who had described themselves as being of Native ancestry. Of these, 42% claimed a Native language as a mother tongue, with 70.6% of those claiming to speak the Native language at home with greater frequency than French or English. Burnaby reported that Native language use was more frequent in northern Ontario than in the southern regions. Watters (in press) reported that standard English is a second language for most NWT children upon entering school. The present author is not aware of current sources revealing the percentages of Inuit who speak only Inuktitut, Inuktitut plus English, or only English.

The practice of language education in the arctic and policies regarding the medium of instruction have shifted with changes in the government bodies responsible for education. Education in NWT was the sole responsibility of the churches, until the federal government assumed control of the school at Tuktoyuktuk in 1948. The federal government enlarged their administrative role until 1969, when responsibility for education was acquired by the territorial government. Whereas the churches had used the Native languages as mediums of instruction in some cases, the federal government instituted English immersion. Preschools were established in some eastern arctic settlements to aid the early learning of English (Wattie, 1968). Students were disciplined for speaking their Native languages at school (King, 1967). Bilingual teaching assistants were hired from within the communities as part of the program for development of English skills (Wattie, 1968). During the period of federal control over NWT education, a Roman orthography was developed for Inuktitut and later modified (Wattie, 1968).

The present educational policy of the territorial government includes a concentration on bilingual education (DIAND, 1981, Government of the Northwest Territories, 1982). DIAND reported the production of workbooks and educational aids in Inuktitut and the Dene languages. The development of instruction in Native languages and the teaching of English as a second language have become priorities for the territorial Department of Education (Watters, in press). Although the language of instruction from Kindergarten to Grade 3 is decided by local education committees, English remains the primary medium of instruction beyond the third grade in NWT schools (Watters, in press).

The assessment battery proposed for use in NWT by Watters (1980, in press) included reading and perceptual tests which used Inuktitut words and text as test materials. The use of both English and Inuktitut materials should allow a more comprehensive and valid assessment of the children's academic and perceptual skills than has usually been the case for Native American children. However, the inferences which can be taken from the test scores of bilingual children will remain more limited than the inferences possible for unilingual children. Some cautions and recommendation regarding the assessment of bilingual children are provided below.

Recommendations for Assessing Bilingual Children

In view of the complexity of relationships among fluency in the first and second languages, the settings in which each language is used, the age at which new languages are introduced, the status of the languages in the society, and achievement on academic and psychological tests, inferences about the cognitive capacity or preferred cognitive strategies of bilingual children from scores on intelligence tests appear to be tenuous at best. Jensen (1980) concluded from his review of the literature pertaining to test bias that IQ tests administered in

English were as valid for the short-range prediction of academic achievement for bilingual children as for unilingual English-speaking children. He warned, however, that scores from such tests should never be interpreted as predictive of long-term achievement or achievement in any but a typical school setting with English as the language of instruction. Cummins (1982) concurred with the view that psychological tests should be interpreted as measures of current academic functioning, rather than learning potential, although Cummins appears to extend that viewpoint to assessment of all children. Jensen suggested that nonverbal tests and social adjustment criteria be included when academic placement decisions are to be based on the assessment results. He also recommended testing bilingual children in both languages and scoring the test in terms of the number of items which are passed in either language.

Cummins (1982) warned that apparent proficiency in conversational English does not imply a commensurate proficiency in English academic and cognitive tests. Toohey (1982) has also cautioned against dismissing a child's bilingualism as a factor in his test score on the basis of observances of conversational English. She noted that Native children in her northern Ontario sample were able to understand directives from the experimenter and to make requests in English, although their productive English scores lagged several years behind the average scores for a neighboring white sample. Toohey argued that following directives and issuing requests to teachers are a stylized use of language by which children are surrounded at school. Teachers and testers may mistakenly infer from such competence that the child is able to understand more complex English instructions for school exercises or test items.

Olsen and MacArthur (1962) suggested that tests used for the cognitive assessment of bilingual children should load only on *g* and should be unrelated to language background. Although reasonable in principle, the utility of these recommendations is limited by the reliability of methods for determining the influence of factors such as *g* or language background. A test's loading on *g* is largely influenced by the method used to define *g* and by the other tests included in the battery, as discussed in Chapter II.D.

Other Factors Affecting Assessment Results

This review has described studies which discussed the performance of Native children on academic measures and psychoeducational tests as the result of cultural deprivation, leading to poor appreciation of literacy and verbal ability or poor selection of cognitive strategies; as the result of the demands for specific skills, imposed by the ecological systems in which they live; and as the result of conflicts between the Native languages and English. Issues such as bilingualism have been discussed as factors which might positively or negatively affect cognitive development and as factors which might obscure a child's present level of ability. Some additional social and health-related issues which are relevant to education and psychological testing in NWT are briefly examined below.

Hearing Impairment

A high incidence of hearing impairment has been cited as a critical education issue in NWT (Watters, 1980). Otitis media, a middle ear infection, has been observed to have incidence rates in the Baffin Island region of NWT which are several times higher than in Southern Canada (Ling, McCoy, and Levinson, 1969). High incidence rates were not evident for nonInuit children on Baffin Island. In a longitudinal study of Alaskan children, onset of otitis media before the age of 2 years, with resulting hearing deficit, was associated with WISC Verbal IQs which were significantly lower than IQs obtained by children with no episodes of the disease before age 2 years or of children who had experienced episodes of the disease without hearing loss (Kaplan, Fleshman, Bender, Baum, & Clark, 1973). Although the severity of hearing loss was associated with the number of episodes a child had experienced, rather than the age of first onset, low Verbal scores and age-grade retardation were associated with both frequency and early onset of otitis episodes. WISC Performance IQs were not significantly affected by a history of the disease.

Schaefer (1980) linked the occurence of otitis media to the practice of bottle-feeding through health surveys of Inuvik, an NWT regional centre, and the relatively isolated community of Arctic Bay, NWT. Arctic Bay's lower incidence of the disease was attributed to the greater tendency for mothers there to breast-feed. When the feeding variable was controlled, higher rates of otitis media were reported for Arctic Bay, which the authors attributed to the harsher climate of that settlement.

Limited support for an association between occurence of otitis media and academic and language ability has been demonstrated with nonarctic samples. Zinkas , Gottlieb, and Shapiro (1978) found an association between severe hearing loss, due to otitis media, and lower scores on all WISC-R IQs, WRAT-Reading and Spelling for a sample of Tennesse school children. Zinkas et al. concluded that occurrence of the disease during critical periods of language and auditory processing development had delayed and impaired the acquisition of language. Brandes and Ehinger (1981) did not find significant differences between the WRAT scores of Otitis suffers and controls in a British Columbian sample, although score trends favored the latter group. Significant differences were obtained on a series of sound blending and matching tasks, favoring children without otitis-related hearing impairment. The intermittent nature of episodes of the disease hinders assessment of its impact on school performance over the long term or psychological test performance on any given day.

Educational Policy

Bowd (1977) has categorized published articles on Native education according to their philosophical orientation toward the purpose of education for Native children. Remedial approaches view the Native child as culturally disadvantaged and requiring special assistance to adapt to the school system. The reports of Downing et al (1975), Knowles and Boersma (1971), and Mickelson and Galloway (1969, 1973) are examples of the remedial approach. The supplementary approach views the problems of Native eduacation as a function of the school-child interaction, and not simply the handicaps of the child. As an example, Bowd cited MacArthur's (1969, 1975a) suggestion that nonverbal media be adapted for educational purposes to overcome English language difficulties and capitalize on spatial skills, while emphasizing the teaching of academic content and skill in the respective languages. Kleinfeld's (1970b) program for teaching syllogisms through Venn diagrams is one application of this approach. The instrumental perspective views schooling as an integral part of the community, with goals which should be decided by the community. Bowd argued that issues in Native education cannot be solved independently of political and cultural issues. The directions in which educational policy is steered will affect the purposes and validity of psychoeducational assessment in turn.

The policies of the territorial government regarding language were discussed in the section of this report which discussed bilingualism. The report of a government inquiry into various aspects of education in NWT made 49 recommendations, calling for increases in local control of education and local provision of services, the training of more Native teachers and improved orientation of teachers from southern Canada, and increased opportunities for adult education (Government of the Northwest Territories, 1982). Recommendations pertaining to special education have been summarized by Watters (in press) as adhering to principles of equal opportunity, individualized programming, integration of students with special needs into the regular classroom, coordination of service delivery, and local provision of service. Applying Bowd's (1977) categories to an evaluation of the report, it appears to combine elements of supplementary and instrumental philosophies of education. While praising the principles underlying the recommendations, Watters (in press) has criticized the report for superficiality and failure to provide a practical plan for implementing the principles. Watters offers some suggestions, including the establishment of a central, short-term residential assessment facility, where individual programs would be developed and the children's educators trained in their use.

Most of the recommendations in the NWT government report cited above, or those of Watter's (1980, in press) proposals for assessment and remediation services, have not yet been implemented. The territorial and federal governments and NWT Native associations are currently discussing the division of the territories into Nunavet and Denendeh, which would roughly cover the lands now inhabited by the Inuit and Dene, respectively. How such a change would affect educational programs and availability of services remains to be seen. In short, educational policy and administration in NWT appears to be in a state of flux. The educational environment to which children will be expected to adapt may change greatly over a short period of time, and from school to school, across the territories. Assessment and educational planning for individual children will have to account for this environmental fluctuation.

Culture Conflict and Technological Change

Conflict between Native cultures and formal schooling has been offered as one explanation of the low achievement of Native children in school and on psychological tests (Sattler, 1982). Some of the conflicts faced by Native children in schools include: school encouragement of competition, rather than cooperation; more rigid discipline and stress on rule compliance by the schools; learning through instruction vs. observation; and attitudes toward rigid time scheduling (Rohner, 1965; Sattler, 1982). Preston (1964) noted the tendency for Alaskan adults to be quiet and appear timid during psychological testing. Bowd, McDougall, and Yewchuck (1982) warn that aggressive question-asking, extensive eye-contact, and other behaviors considered appropriate for teachers in nonNative schools, may be considered rude by Native children.

The shyness and quietness of Native children in the classroom is often attributed to characteristics of Native culture, as in the articles cited immediately above. These attributes are thought to reflect a mismatch of learning and teaching styles. It is seldom interpreted as children's response to a particular political situation. Kleinfeld (1973a) provided some support for her hypothesis that Native children's classroom participation was a function of perceived threat, employing teacher and self-ratings of participation and an environment-perception questionnaire. King (1967) observed the eagerness with which Yukon Native children approached learning upon entering the residential school which he studied. He noted that they would seek out friendships with the teaching staff, until the response of the latter group demonstrated that acceptance was to be gained through maintenance of order and discipline rather than exploration. After a short time at the school, children would engage in gamesmanship and preferred to avoid making open decisions. The behavior of the staff at that school can not be generalized to teaching staff across the Yukon and NWT. However, King's ethnographic account illustrates the manner in which behavior which is thought to be typically "Native", such as social withdrawl in the classroom, may be a product of the environment created by nonNative authorities. Kleinfeld (1975) has noted that, although Native culture is reflected in the behavior of Native children, all such behavior is not necessarily a manifestation of Native culture. This argument was stated in the context of her observations of the tendency for researchers and teachers to attribute all behaviors of their Native students to "Native culture".

Cultural conflict has been identified as one factor underlying the high rate of school-leaving in the Arctic. Native Parents have expressed ambivalence about the schools, which they perceive to be a means to mobility in white society (Kleinfeld, 1971a). While they hope that formal education will lead to good jobs for their children, they recognize the probability that such jobs will often mean separation from the home community. Kleinfeld reported that some parents expressed relief when their children dropped out of school, since this would reunite the children with their families. This result does not always occur, however. While some children leave school to help at home or on hunting trips, other dropouts simply wander around their communities (Gall, 1980; Watters, 1982).

School attendance is often sporadic for those children who do remain. Watters (in press) cited family hunting, fishing,and trapping activities, which may conflict with the school term, as one factor drawing children out of school. During the summers, the long daylight hours disrupt schedules, as children often play or work at home during the early morning hours and may sleep through the school day. Watters reports that surveys of children in the Kitikmeot region, who are typically achieving below the grade level expected for their age, have indicated that their school achievement is at the level expected, given their actual attendance figures.

The introduction of television into several arctic communities in the last decade appears likely to affect attitudes toward, and readiness for, education. Watson (1980) observed that children in Rankin Inlet spent more time at home and were less interested in joining hunting expeditions after the introduction of television. He also observed that residents became more schedule-oriented in response to television, and that fishnets were checked less regularly during popular programs. Teachers and parents reported that children were learning English at an earlier age. Some adults watched programs such as Sesame Street as an aid to learning English. The introduction of television to Frobisher Bay, Baffin Island, was linked to an increase in awareness of international events, consideration of a broader range of occupations for their children, and increased interest in travel outside of the local region (O'Connell, 1975). Both O'Connell and Watson reported that action shows were preferred over documentaries or interviews, as many of the residents of the respective communities did not speak English with sufficient fluency to understand the latter. Inuktitut programming is now produced in larger centres, such as Rankin Inlet (Devine, 1981). The effect of such programming on educational readiness and achievement is yet to be assessed.

In earlier discussions of the field independence research of Berry and MacArthur, the relative field independence of Inuit subjects was attributed to their nomadic, hunting lifestyle, their relatively permissive child-rearing practices, and the loose organization of their communities, with emphasis on personal autonomy. Vallee (1968) has observed a trend toward development of class structure in the settlements of the Keewatin region. Inuit living in the settlements on a permanent basis tended to be wealthier than those who lived off the land. Inuit settlement leaders formed a group which cut across settlements. Vallee noted an increased division of labor among settlement Inuit, with status varying among the respective occupations. Vallee conceded that the Inuit living in smaller camps, who retained the traditional loosely-structured societies, did not necessarily recognize the settlement Inuit leaders as such, or view themselves as lower status. The pervasiveness of this shift away from the societal organization and livelihood traditionally practised by the Inuit may affect opportunities to acquire the spatial and mechanical skills attributed to them. MacArthur (1968) warned that rapid economic development in northern Canada forces construct validity research to be an ongoing operation.

Attempts to Reduce Bias in Assessment

A catalogue of methods for reducing culture bias in psychological tests was presented in Chapter II.A of this report. The method recommended by any one author reflects the authors definition of test bias. The recommendations for reducing bias in pyschological assessment of Native North Americans tend to reflect the viewpoint that the presence of bias may be determined by group differences in mean scores on the test. More and Oldridge (1980) noted that Natives are overrepresented among those labelled as EMR in British Columbia. They have argued for the standardization of the SOMPA battery (Mercer, 1979) for Natives in that province. They also state that the use of psychological tests as tracking devices is unethical for such children, as this procedure denies a full education to those labelled as handicapped.

MacArthur's assessment studies with Alberta, NWT, and Greenland samples appear to comprise the most extensive program of research into the psychometric properties of psychological tests for use with Native Canadian children. His criteria for assessing the construct validity of a test to be used as a measure of general intellectual potential are: sampling of a broad factor of intellectual tasks; relatively small mean differences between cultural groups; a moderate relationship with current school achievement and ability to predict future performance on intellectual tasks; internal consistency and stability; little reliance on test sophistication, with easily understood instructions which do not rely heavily on language. He recommends practice with similar types of items, and discourages the use of speeded tests (MacArthur, 1968). When two or more cultural groups are being compared, the test stimuli should be equally familiar to all groups (West & MacArthur, 1964). Using these criteria, MacArthur selected a battery of tests he considered to be culturally reduced, i.e., less biased than most of the language-dominated tests available for intellectual assessment. This battery included Raven's CPM and SPM, Lorge Thorndike Nonverbal subtests, and other spatial and achievement tests. Norms were provided for Indian and Inuit samples from NWT and the Yukon Territories (MacArthur, 1965). MacArthur appears to have placed the greatest confidence in Raven's SPM, which he recommended for identification of talented students from areas with limited formal schooling, for students with foreign language backgrounds, or children who are otherwise handicapped by convential verbal intelligence tests (Elley & MacArthur, 1962). The strength of Raven's Matrices lay in its arrangement of items, which MacArthur (1968) suggested form an age scale, beginning with tasks demanding simple perceptual skill and progressing through reversible concrete operations to formal operations tasks. The item arrangement within each of the five sets of SPM also allow a sample of the child's ability to learn the task demands from the earlier items and apply this knowledge to more difficult items.

Hynd and Garcia (1979) have argued that tests such as the WISC-R are still useful as measures of acculturation and present functioning in English. However, they recommended that nonverbal measures be used to assess mental handicaps and that criterion-referenced tests be included test batteries for use with Native children. They also recommended modeling of test-taking strategies in the classroom and before administration of the tests. Like MacArthur's (1968) recommendation of practice, this procedure is intended to increase the child's test-wiseness.

Summary

The WISC and WISC-R profiles of Native North American children have typically consisted of Verbal Scale scores which are more than one standard deviation below the U.S. mean, and Performance Scale scores within the average range. This trend has also been evident in the scores of Native samples on other test batteries which included tests with both verbal and nonverbal response demands. Several samples of Native children have exceeded U.S. norms on various nonverbal tests, such as the Draw-A-Man.

Valdity indices for the Wechsler tests and other batteries have indicated that these tests do not measure the same constructs for children in some Native communities as they have been assumed to measure for nonNatives. The isolation of tests such as Raven's SPM from other nonverbal tests, in a factor labelled by MacArthur as Inductive Reasoning from Nonverbal Stimuli, led to the conclusion that the verbal and nonverbal ability dichotomy was a simplistic explanation for the academic difficulties and low verbal test scores of Native children. The field independence research of Berry and MacArthur provided a potential theoretical rationale for variance in nonverbal scores across Native samples

The test score patterns of Native samples have led some some investigators to make inferences about the capacity of Native children for language use or for verbal mediation of behavior; about their ability to process information sequentially; or about their preferred cognitive strategies for academic tasks. In most cases, these inferences were based on studies with experimental designs which were inadequate for the questions addressed. Some of the inadequacies were identified in the above review. Among the most serious flaws in much of the literature is the tendency to make inferences about language capacity on the basis of performance in the child's second language. Even when children were more fluent in English than in the Native language, properties of the latter, such as the irrelevance of voicing for distinguishing consonant sounds in Cree, may have interfered with test performance if the children's comprehension and production of English was influenced by adults who spoke English as a second language.

Research on the effect of bilingualism on psychological test scores indicates that the relationship is affected by a complex array of variables pertaining to the status and common uses of each of the languages spoken. Research on the second language learning of minority children has tended to support the hypothesis that immersing a child in a second language, with no opportunity to continue the development of the first language, leaves the child with inadequate abilities in both languages. The network of languages and dialects spoken across the Northwest Territories is sufficiently complex to defy global statements about the impact of biligualism on academic achievement. However, the many facets of bilingualism cannot be discounted as confounding variables in studies which have led to inferences about verbal capacity.

Recommendations for the psychological assessment of Native children have followed principles similar to those underlying recommendations for the testing of all bilingual children. Tests of cognitive ability should not require verbal sophistication to respond or to understand instructions. Tests with extensive language use should only be considered as short-term predictors of success in the conventional North American classroom setting.

Watters (in press) has reported that tests thought to be culture-fair have proven to be less useful for assessment with NWT children than has the WISC-R, Bender Motor-Visual Gestalt Test, or Draw-A-Man. She did not specify the shortcomings she attributes to tests such as the Raven SPM and CPM. The clinical utility of the tests chosen by Watters for use with NWT children remains an empirical question. The present study addresses one aspect of the question for the children in the Keewatin and Kitikmeot regions. Specifically, the study asks whether the ability factors attributed to the WISC-R, and assumed in the interpretation of test profiles for U.S. children, can be identified from the correlations among the subtest scores of school-children in the Keewatin and Kitikmeot regions. The following section discusses some methodological issues pertaining to the testing of factor analytic models across populations.

D. Testing Factor Models

The factor analytic studies of cognitive ability discussed in this chapter have employed a wide variety of factor extraction and rotation techniques. The close association of factoring methods to intelligence theories led Sternberg (1980, 1981) to suggest that exploratory factor analysis is not well-suited to confirming or disconfirming theories of intelligence. This section ~~of~~ Chapter II explores methodological weaknesses in the theory or common practice of factor analysis that limit the reliability of its results. The implications of the indeterminacy of factor scores for factor theories are noted. Common flaws detected in published studies are briefly listed and disagreements regarding methodology, some of which were mentioned in the section on factor theories of intelligence, are summarized. Confirmatory factor analysis and factor matching procedures are introduced and their power to increase confidence in factor theories is examined. The final and major part of this section will focus on procedures for confirmatory factor analysis using maximum likelihood methods.

Reliability and Validity of Factor Analysis Results

In earlier discussion of the application of factor analytic solutions for intelligence test batteries, the validity of computing factor scores, such as a Verbal Comprehension score from the WISC-R, was examined. Computational formulae were cited for the three WISC-R factors. However, such factor scores are necessarily only estimates of a person's score on the theoretical construct presumed to underly the observed variables. Since the multiple correlation between scores on a common or specific factor and the observed variables is not unity, if only because the latter scores include measurement error, the factor scores are not uniquely determined by the observed variables (McDonald & Mulaik, 1979). In other words, scores on the Verbal Comprehension factor cannot be precisely calculated by applying the loadings to the subtest scores. Factor score indeterminacy has implications for theory-generation from exploratory factor analysis as for well as the validity of computing factor scores. This indeterminacy

> does not merely concern the problem of obtaining a score, or the problem of obtaining too many scores, that might be the score of an individual on a common factor. Rather it concerns the inability of a finite set of observed variables in an exploratory factor analysis to determine unambiguously what attribute of the individual the factor variable represents (McDonald & Mulaik, 1979, p. 299).

As Jöreskog and Sörbom (1978) point out, a given covariance matrix may be represented by several sets of factor pattern, factor intercorrelation, and unique variance matrices. Therefore, the factor pattern presented in support of a factor theory is only one of several possible such patterns. The difficulty of testing a factor model becomes apparent in the face of the large array of factor solutions obtainable by rotation from a given covariance matrix.

A factor structure which supports a theoretical factor model may also be an artifact of the use of a restricted group of observed variables. The factor structure of all tests within a behavioral domain may be inconsistent with the factor structure of a subset of tests within that domain. Even the use of marker variables, i.e. variables which have been demonstrated to load on a common factor and interpreted as a particular construct, does not assure the researcher that the factors reflected in the common variance for the marker variables are identical in the context of two different test batteries (McDonald & Mulaik, 1979). Ascription of specific psychological properties to factors produced through exploratory factor analysis of a limited number of variables is therefore necessarily tentative.

Although the implications of factor score indeterminacy listed above are central to the validity of the interpretation of factor patterns, the literature on indeterminacy is not reviewed in detail in this report. McDonald & Mulaik (1979) provide a relatively nontechnical discussion of the topic, while Schönemann (1981) critiques the history of factor analytic research on intelligence in the context of such issues as indeterminacy. Some of the other methodological problems plaguing factor analytic research are mentioned below.

Comparison of Guilford's (1952) list of common faults in factor analytic research with Comrez's (1978) list suggests that much has not changed in the practice of such research in the last 30 years. Faults common to both lists include the use of noncontinuous or nonnormally distributed variables, extraction of too many factors for the number of variables, and analysis of spurious correlations. Harman (1967) notes that that many researchers interpret the factor structure matrix (the correlations of variables with factors) rather than the factor pattern matrix (the coefficients of the factors in the factor model expression of the variables). Harman suggests interpretation of the latter, as it indicates the direct contribution of the factor to each variable, but notes that the term "factor loading" is often used ambiguously in published reports of oblique solutions, making it difficult to determine which matrix is being discussed. Nunnally's (1978) use of the term "loading" to denote elements in the factor structure matrix, in contrast to Harman's use of the term in reference to the pattern matrix, is one example of the extra confusion created by communication difficulties in the literature.

The method of factor extraction is also a source of controversy which has implications for cross-cultural cognitive research. The discrepancies in the solutions obtained by Jensen and Carroll for a set of reaction time data, which were reviewed in Chapter II.A, provide one example of the differences in results that arise from variations in application of the factor analytic model. Jensen's (1979, 1980) conclusions regarding cultural differences on cognitive tests, as a function of their *g* loadings, were not supported when his data was subjected to hierarchical analysis, as Carroll (1981a, 1981b) found no evidence for a general factor. Carroll (1981b) dismisses the interpretation of principal components as he claims that these components are inflated by unique variance. The charge that any factor theory may be supported by selection of the proper extraction and rotational procedures appears to gain support from such disagreements.

The use of criteria such as simple structure for factor rotation is presented by some researchers as sufficient precaution to allow theoretical interpretation of exploratory factor analytic results. Carroll (1980) responded to Sternberg's (1980) charges with the claim:

> With well-designed studies, the principles of simple structure can pretty well dictate the final solution. Parsimony is the essential principle underlying the idea of simple structure; it says that one wants to account for a given variable with the minimum number of factors - often with only one factor (p. 15).

This position is contrasted with Comrez's (1978) contention that simple structure seldom fits the variables used, i.e. that psychological variables are generally factorially complex. Comrez adds that allowing high intercorrelations among factors to achieve simple structure results in factor interpretation problems, though of a different sort than those involving complex variables.

In summary, the factor analytic procedures practiced and endorsed do not reflect consensus among the leading researchers in the field. The interpretation of factor analytic results must include recognition of the problem of factor score indeterminacy, and the degree to which the results may be artifactual. The replication of factor studies is also subject to these methodological concerns. The next section of Chapter II explores various methods of confirmatory factor analysis and the increased generalizability these methods provide to factor analytic results.

Factorial Invariance and Confirmatory Factor Analysis

The extent to which factors thought to underlie a set of observed variables may be considered fundamental constructs depends on the generalizability of the factor definitions to varying conditions. If different sets of variables in the same behavior domain are factored, or a set of variables is analyzed across several populations, with replication of the factors, this is a demonstration of factorial invariance (Mulaik, 1972). The comparative factor analytic studies reviewed earlier in this chapter were generally concerned with demonstrating the invariance of factors of mental abilities across populations.

Confirmatory factor analysis involves the test of the accuracy of an a priori factor model for a particular data set (Mulaik, 1972). The latter model may be derived from theory or from the factor pattern of another sample. Mulaik appears to discuss factorial invariance and confirmatory factor analysis as separate, if related, issues. However, these issues are discussed jointly in the present paper. The demonstration of construct validity of a test battery for an arctic population was defined in Chapters I and II as the question of generalizability to that population of the factor structures implicit in the interpretation of the test. This is a question of factor indeterminacy across populations. The factor structures in question have been hypothesized a priori, based on the factor analytic studies of the WISC-R reviewed in Chapter II.B. The methods of confirmatory factor analysis chosen for the present study allow the a priori factor model to be tested against the factors obtained from the data. Other methods allow the researcher to study factor invariance across populations without specifying the factor model before the study begins (although Mulaik warns that the variables should be carefully chosen on theoretical grounds).

Confirmatory methods differ in their power to test various aspects of a given model, e.g. while some methods allow tests on the number of factors or the values of the factor intercorrelations, other methods allow examination of only the pattern matrix. A number of confirmatory methods are described below. The method labelled Procrustes rotation is described in some detail as this method is frequently cited where factor patterns are compared across samples.

Procrustes Rotation to Congruence

Meredith (1964) argued that when two covariance matrices are sampled from a single population and independently factor analyzed, one would not expect invariant factor patterns to emerge. However, rotation of one factor matrix to agree with the other should result in nearly invariant factor patterns in such a case. Rotation of the pattern matrix is achieved through multiplication with a transformation matrix θ, which is derived for each of $i=1,2,...k$ pattern matrices Λ_i so that

$$\Lambda_i \theta_i^{-1} = \beta \qquad \text{(II.3)}$$

where β is a common target matrix. The k transformation matrices are calculated to minimize

$$\phi = \sum_{i=1}^{k} \mathrm{tr}[e_i' e_i] \qquad \text{(II.4)}$$

where e is a matrix of differences between the loadings on the rotated matrix for sample i and those of the target matrix. For example, if one wishes to rotate two sample pattern matrices to maximum congruence, or similarity, the derived transformation matrices would rotate the respective pattern matrices to maximum similarity with a common target matrix. The fit of a rotated matrix to the target would be measured by calculating differences between their corresponding loadings. The matrix of these differences, or errors, is named "e". The diagonal elements of e'e are equal to the sum of squares of errors across loadings for each factor and the trace is equal to the sum of these terms across all factors. Dividing the trace by the number of factors in the pattern matrix gives the average error sum of squares for the factors for that sample and this value may be used as a goodness-of-fit measure. Minimizing the sum of tr[e'e] across samples therefore minimizes the total sum of squares for fitting factors across samples. Meredith (1964) stated that the transformed matrix may be oblique.

A solution for orthogonal rotation of a sample matrix to a target matrix is provided by Schönemann (1966). Like Meredith's solution for oblique rotation, this method operates by deriving a transformation matrix that minimizes tr[e'e] when e is equal to the difference of the rotated sample factor matrix and the target matrix. Schönemann referred to this factor-matching rotation problem as the orthogonal Procrustes problem. A comparable method for orthogonal rotation of two factor matrices to congruence rotates the two matrices so as to maximize

$$\theta = \sum_{n=1}^{m} \sum_{j=1}^{j} P_{jn1} P_{jn2} \qquad \text{(II.5)}$$

where P1 and P2 are the rotated JxM factor patterns for samples 1 and 2, respectively (Cliff, 1966). This value is equal to the sum of cross-products of loadings on corresponding factors from the two rotated factor patterns. It is also equal to the trace of the matrix product of the two rotated matrices.

Given the difficulties with establishing the reliability and uniqueness of factor solutions, rotation to a hypothesized target matrix, using the methods described by Meredith (1964), Schönemann (1966), and Cliff (1966), has been proposed as a confirmatory approach to factor analysis with single-sample data (Armstrong & Soelberg, 1968). Horn (1967) has challeged the scientific value of such methods, however, labelling them as subjective. Whereas objective rotation procedures would produce the same results for researchers with different hypotheses, subjective measures allow the rotation of the pattern matrix to be affected by the researcher's theory about the results.

Horn's (1967) concern's about the power of subjective rotation procedures to support any theory tested were substantiated in a series of simulation studies. Horn generated scores on 74 random variables and named these variables after tests which he had used in earlier research. A hypothesisized factor pattern matrix was designed on the basis of previous research with the actual tests and the principal factor solution for the 74 variables was rotated to maximum congruence with this target matrix. Although Horn's criteria for accepting a loading as salent was .20 and salient loadings tended to be below .30, Horn concluded that, since 54 of the 65 hypothesized loadings were indeed salient, seemingly interpretable results could be obtained from subjective rotation of random data. Horn and Knapp (1973) reversed the procedure used in Horn's (1967) experiment, rotating factor patterns from real data to randomly-generated target matrices by orthogonal Procrustes. The data was taken from three studies reported by Guilford and Hoepfner (1971) in support of Guilford's Structure of Intellect (SI) model of intelligence. Using a salience criteria of .30, Horn and Knapp were able to confirm as many hypothesized loadings on the randomly-generated model as Guilford and Hoepfner had confirmed on the SI model. The power of subjective rotation procedures to support factor theories with random data and random thories with real data undermined the value of such procedures for confirmatory analysis.

Guilford's (1974, 1977) response to these challenges was to assert that the replication of his hypothesized factor models in thirty studies was evidence of their reliability and validity. Horn and Knapp (1973) countered this argument by stating that 30 applications of a procedure which could be shown to support nonsubstantive theories were no more impressive than a single application of such a method. Guilford (1977) also challenged the accuracy of objective rotation methods, citing the results of analyses of correlation matrices derived from a priori factor models with simple structure. When varimax and promax rotations were performed on the factor patterns for these correlation matrices, discrepancies from the original factor patterns usually took the form of additional salient loadings for variables. Guilford concluded that adherence to simple structure was a measure of accuracy for analytic rotation procedures such as varimax and promax. The weakness in this reasoning appears to lie in its failure to account for the fact that a given correlation matrix can be represented by several factor patterns. The fact that Guilford derived the correlation matrices in this study from his theoretical model does not necessarily imply that a discrepant factor model could not be equally descriptive of the correlation matrix. The threats to the validity of subjective rotation procedures which Horn and Knapp raised are not eliminated by Guilford's counterarguments.

The weaknesses of the Procrustes procedures may be limited to their application to factor matrices with characteristics which would be a source of concern in any type of analysis. Horn and Knapp (1974) admitted that they were not able to rotate data to a nonsensical target matrix with simple structure if the ratio of variables to factors was 5 or greater. Simulation studies which controlled for various characteristics of the factor matrices indicated that the vulnerability of Procrustes rotation to capitalization on chance is a function of sample size, the number of observed variables, and the ratio of the number of variables to the number of factors (Humphreys, Ilgen, McGrath, & Montanelli, 1969; Nesselroade & Baltes, 1970; Nesselroade, Baltes & Labouvie, 1971). Humphreys et al. rotated randomly-generated factor matrices to arbitrary target matrices by orthogonal and oblique Procrustes. The best fits to these arbitrary factor patterns was achieved with small samples, relatively numerous variables, and/or low variable-to-factor ratios. The strongest effect was attributed to the number of variables per factor, with a ratio of 2.0 resulting in closer fits than a ratio of 4.0. Nesselroade and Baltes (1970) found that factor matrices from random data could be orthogonally rotated to similarity with greater success when the number of variables was low and/or the number of factors was large. Sample size was not consistently related to the similarity between rotated factors matrices and the interaction between the number of variables and the number of factors was not consistently significant. These results were replicated with oblique rotation to similarity (Nesselroade et al., 1971). The implications of these successes in fitting factor models with random data are limited by Pennel's (1972) finding that large factor loadings in Humphrey et al.'s factor matrices were found to be nonsignificant when confidence intervals were calculated.

The practice of confirmatory factor analysis through Procrustes rotation of the obtained factor matrix to maximum congruence with a target matrix has been demonstrated to be capable of capitalizing on chance in producing theoretically "meaningful" results with nonsense data or factor models. Nunnally (1978) has recommended that factor analysts refrain from using Procrustes rotations for this reason, arguing that these methods are still in the exploratory stage. In attempts to allow researchers to test factor theories with more confidence than Procrustes rotation appears to allow, a number of indices of a factor pattern's fit to a theoretical model have been derived. Some of the more prominent of these are described below.

Indices of Factor Similarity

The congruence coefficient (Wrigley & Neuhaus, 1955) is a measure of similarity of the loadings on two factors. Its purpose in confirmatory factor analysis or factor matching studies is to compare factors across two distinct factor pattern matrices. This coefficient is calculated as

$$rc = \sum_{i=1}^{n} b_{i1} b_{i2} \bigg/ \left(\sum_{i=1}^{n} b_{i1}^2 \sum_{i=1}^{n} b_{i2}^2 \right)^{1/2} \qquad \text{(II.6)}$$

where b_{i1} and b_{i2} are the loadings of variable i on the two factors to be compared. This coefficient is similar in form to a correlation coefficient, ranging from -1.0 to 1.0, but it differs in that the former index considers loadings as deviations around a value of 0 rather than deviations around the mean loading for a given factor. Since the size and sign of loadings is of more interpretive value than their distance from the factor's mean loading, the congruence coefficient is a better index of factor similarity than the correlation coefficient (Cattell, 1978). Comparison of two factor pattern matrices involves construction of a matrix of congruence coefficients such that the element in the first row and second column is the congruence coefficient for Factor I from the first pattern matrix and Factor II from the second. If the pattern matrices are similar, the diagonal elements of the matrix of congruence coefficients should be very high while the off-diagonal elements should be close to 0.

The congruence coefficients are usually calculated after Procrustes rotation. The formula for the coefficient is closely related to the criteria for Cliff's (1966) formula for rotation to congruence. In fact, Cliff's procedure effectively maximizes the congruence coefficients. Concerns regarding the reliability of results from Procrustes rotations also restrict the interpretability of congruence coefficients.

One major fault of the congruence coefficient is the lack of a clear test of significance based on a theoretical sampling distribution. However, Cattell (1978) and Korth

(1978) provide critical values of *rc* which were derived from Monte Carlo experiments for a limited number of variables and factors. Skakun (1971) provides critical values for a related statistic, tr[e'e], derived from a series of Monte Carlo studies with factor matrices of various orders. This statistic was minimized in Schönemann's (1966) procedure for orthogonal Procrustes rotations.

A nonparametric test of the similarity of two factors was derived which focuses on the decision to accept a loading as salient rather than the relative or absolute size of the loading (Cattell & Baggaley, 1960; Cattell, Balcar, Horn, & Nesselroade, 1969). The salient variable similarity index, *s*, is calculated by categorizing the loadings on each factor as positive salient, nonsalient, or negative salient, as determined by an a priori critical value. The categories of the loadings for each variable are cross-tabulated for the two factors. If the factors are identical, only the diagonal cells of this table will have frequencies greater than 0, i.e., variables which had positive salient loadings on the first factor will have positive salient loadings on the second factor, etc. Maximum similarity of factors is indicated by an *s* of 1.0; maximum dissimilarity, by an *s* of -1.0. Maximum dissimilarity merely indicates that the factor needs to be reflected. An *s* of 0 indicates that the frequency of loadings which match across factors is not greater than would be expected by chance. Formulas for calculating the expected frequency of each cell and tables of critical values of s are available in Cattell et al. (1969) and Cattell (1978). Cattell (1978) recommends the use of s in combination with the congruence coefficient to test the similarity of factors across samples. Although s is a less powerful statistic than the parametric indices of factor similarity, it avoids the assumption that the loadings are normally distributed on a ratio scale (Cattell & Baggaley, 1960).

Nunnally (1978) describes separate procedures for the matching of factors across samples and the fitting of sample factor matrices to a theoretical model. In the former case, Nunnally suggests calculating two sets of factor scores for the subjects in each sample. These scores would be derived using the factor loadings from the the two sample factor matrices. Correlations of factor scores are calculated within each sample. Scores on a given factor as calculated by the two weighting systems would be expected to have a high correlation if the factor patterns were similar across groups.

Nunnally (1978) suggests multiple group factor analysis for the testing of factor models to sample correlation coefficients. The hypothesized factor matrix contains values of 1.0 for loadings which are expected to be salient and values of 0 for all other loadings. Each variable loads on only one factor. The correlation of a variable to a factor is calculated by applying the variable intercorrelations to the formula for calculating a variable's correlation with the sum of a set of variables. A factor model is disconfirmed when a variable assigned to a group has a low correlation with the factor for that group and/or has a large correlation with the factor for another group. The matrix of residual correlations will contain some large values if the factor model is not supported. This confirmatory procedure is inappropriate where there is a small number of variables in each group, as each variable will be prominently represented in the factor score, thus inflating the variable-factor correlation.

One disadvantage of all the indices discussed thus far is their neglect of the correlations among factors. The confactor method of rotation attempts to account for the the correlations among factors as well as the variables' loadings on the factors in fitting two factor structures into the same factor space (Kaiser, Hunka, & Bianchini, 1969). The fit of the two factor patterns is measured by the cosines between all pairs of corresponding factors from the two samples. The confactor method of rotation is an improvement upon the other indices in this regard. The authors warn that an element of judgement is required in applying the method.

The applications of maximum likelihood factor analysis (MLFA) are discussed below. Many of the weaknesses of the rotation procedures and similarity indices described above are contrasted with the detailed information regarding a model's validity provided by ML analysis. The weaknesses and assumptions of this method are discussed, and its suitability for the present research examined.

Maximum Likelihood Methods

The application of maximum likelihood factor analysis to confirming factor models and matching factors across groups has been developed largely through the work of Jöreskog (1969, 1971, 1978; Jöreskog & Lawley, 1968; Jöreskog & Sörbom 1978, 1981). Although the mathematical theory of likelihood functions had been developed before this time, the method was not practical with the limited power and efficiency of early computers. The analytical capabilities of Jöreskog's earlier programs have been combined in **LISREL** (Linear Structural Relationships) (Jöreskog & Sörbom, 1978, 1981). A theoretical treatment of MLFA is available in Lawley and Maxwell (1971). The hypotheses which may be tested with MLFA are examined below, followed by a discussion of the assumptions associated with this statistical model and the robustness of the statistics to violations of these assumptions. The procedures of hypothesis testing with Jöreskog's programs, particularly **LISREL**, are then described and implications of deviations from these procedures in several published studies are examined. Finally, MLFA is compared to the other confirmatory methods and associated statistics described in this chapter, leading to the rationale for applying **LISREL** in the present study.

Hypothesis Testing

The factor analysis model of the components underlying the covariance matrix for a set of observed variables is expressed algebraically as

$$\Sigma = \Lambda\Phi\Lambda' + \Psi \qquad \text{(II.7)}$$

where Σ is the variable covariance matrix, Λ is the factor pattern matrix, Φ is the factor intercorrelation matrix, and Ψ is the covariance matrix for the uniqueness, or error, terms. In ML analysis the parameters within the matrices on the right side of the above equation may be free to be estimated, fixed to a specific value, or constrained to be equal to the estimated value of another parameter. An iterative procedure is employed in which values for Σ and the free parameters of Λ, Φ, and Ψ are estimated and compared to the obtained sample covariance or correlation matrix, S, until the best possible fit between S and an estimate of Σ is obtained. The likelihood function

$$\log L = -\tfrac{1}{2}N[\log|\Sigma| + \mathrm{tr}(S\Sigma^{-1})] \qquad \text{(II.8)}$$

where N is the sample size associated with S, is maximized by the above procedure. This is equivalent to minimizing the loss function

$$F = \log|\Sigma| + \mathrm{tr}(S\Sigma^{-1}) - \log|S| - p \qquad \text{(II.9)}$$

where p is the number of observed variables. The loss function is more convenient to calculate and is therefore used in LISREL. Multiplying F by the sample size, N, gives a value equal to $-2(\log L)$ which has a χ^2 distribution with

$$d.f. = \tfrac{1}{2}(p)(p + 1) - t \qquad \text{(II.10)}$$

degrees of freedom, where p is the number of variables and t is the number of parameters left free to be estimated by the iterative process (Jöreskog, 1971; Jöreskog & Sörbom, 1978).

The null hypothesis tested by this χ^2 is that S is explained by the model defined by the restrictions placed upon Λ, Φ, and Ψ by the researcher to simulate a theoretical factor model. This appears to be equivalent to stating H_0: $\Sigma = S$. The alternative hypothesis, H_1, is that Σ is any positive definite matrix. Bentler and Bonett (1980) refer to this H_1 as the satiated model. Specification of the parameters in the restricted model may be ordered to allow tests on the number of factors, the fit of a specific factor pattern, or the similarity of a subset of parameters across samples. Standard errors may be calculated for parameter estimates, assuming a normal distribution for estimates whose absolute value does not exceed 0.6 (Lawley & Maxwell, 1971).

The null hypothesis may also be tested against a restricted model in which one or more of the parameters which were fixed for H_0 are free for H_1. Where F_0 is the minimum value of F under H_0 and F_1 is the minimum value of F under H_1, $N(F_0 - F_1)$ has a χ^2 distribution with $t_0 - t_1$ degrees of freedom (Jöreskog & Sörbom, 1978). This test can be performed more directly by calculating the difference in χ^2 values obtained by the two models against the satiated model and interpreting this value as a χ^2 (Hays, 1964). Where the free and constrained parameters in the factor model are defined as members of a vector of parameters, Θ, the comparison of two models may be considered to be a test of the hypothesis H_0: $\Theta_0=\Theta_1$. This null hypothesis implies that all parameters fixed to 0 in the more restrictive model will equal 0 if estimated in the less restrictive model. The procedures for testing the incremental fit of sequential models is described in more detail later in this chapter.

At first glance the array of statistical tests available with MLFA appears to resolve the problems attending such subjective methods as Procrustes rotation. In practice, however, MLFA is not an entirely objective procedure and judgement is required by the researcher at various steps in the analysis. Hypothesis testing is an inherently judgemental process which allows a decision or risk evaluation, rather than assertions (Bakan, 1966). Jöreskog's advice on the interpretation of results of statistical tests with MLFA reflect Bakan's contention that the null hypothesis in social science research is generally false, (e.g., that population means, proportions, etc. are almost never identical across groups). Jöreskog (1978) suggests that the statistical problem for MLFA is that of fitting various models and deciding when to stop fitting, rather than testing a given hypothesis "which a priori may be considered false" (p. 448). The statistical problem is further defined as the task of extracting as much information as possible from a sample without "going so far that the result is affected to a large extent by noise" (Jöreskog, 1978, p. 448). For instance, the decision to accept two factors as sufficient to explain a given covariance matrix on the basis of a nonsignificant χ^2 does not assure the researcher that there are no additional factors. Rather, it indicates that there would be no point in fitting further factors to the data as these would be indistinguishable from sampling error (Lawley & Maxwell, 1971; Jöreskog & Lawley, 1968). As a further qualification to the interpretability of statistical tests on specific factor models, Lawley and Maxwell (1971) argue that the factor analytic model in general is "like other models, useful only as an approximation to reality, and it should not be taken too seriously" (p. 38).

One consequence of the above limitations on the interpretation of statistical test results for confirmatory factor analysis is that the question examined is not "Is the clinical factor model, e.g. Kaufman's three-factor model, a true representation of the relationship among WISC-R subtests?" The question becomes "Does the clinical model sufficiently explain the relationship among subtests so that modifications to that model would not result in improvements to the explanation which would be distinguishable from sample fluctuation?" The legitimacy of applying MLFA to even this less ambitious question is dependent upon the data's adherence to the assumptions associated with the likelihood distribution. These assumptions and their importance are discussed below.

Assumptions of Statistical Tests

Jöreskog and Sörbom (1981) list three assumptions for the use of the χ^2 test for model-fitting. These are:

1. All observed variables have a multinormal distribution.
2. Analysis is based on the covariance matrix rather than the correlation matrix.
3. The sample size is large.

Jöreskog and Sörbom note that these conditions are rarely simultaneously satisfied in practice and recommend that decisions be based on the relative size of χ^2 to the degrees of freedom. Bentler and Bonett (1980) contend that little is known about the robustness of the χ^2 test for typical applications in model-fitting. The necessity of the second and third assumptions is not uncontested in the literature. Contrasting views and evidence regarding those two assumptions are discussed below.

Analysis of the Correlation Matrix

Jöreskog (1971) argued that analysis of the correlation matrix violates the assumptions of the likelihood ratio statistic but does not give reasons why a linear transformation of raw scores to *z*-scores would affect inferences made regarding the pattern of proportions of shared variance among subtests. The argument is repeated by Werts, Rock, Linn, and Jöreskog (1976). Lawley and Maxwell (1971) suggest that, while analysis of correlation matrices is a common practice in educational and psychological research, this practice causes problems when the distributional properties of estimates are of interest or when significance tests are required. Lawley and Maxwell contend that the variances and covariances of the latent roots of a correlation matrix are more complex in form than the corresponding properties of a covariance matrix. These authors prefaced an example of ML analysis of a correlation matrix with the disclaimer that provision of the example did not constitute endorsement of the practice (Lawley & Maxwell, 1971).

Lawley and Maxwell (1971) appear to be inconsistent in their assessment of the legitimacy of analyzing the correlation matrix. In contrast to the cautions stated above, they claim

> A satisfactory feature of the maximum likelihood method is that it is independent of the scales of measurement of the variates.... This is of great computational convenience, since it means that in performing numerical calculations we can standardise all variates and substitute for S the sample correlation matrix of X, having unities as its diagonal elements. (Lawley & Maxwell, 1971, p. 33)

This argument is reinforced by the later statement:

> Let us henceforth assume, as is usually the case in practise, that all the Ψ_i are free parameters and that all fixed parameters in Λ are zeros. It is then

> easy to verify that the method of estimation of this chapter is invariant under changes of scale of the x-variates. (Lawley & Maxwell, 1971, p.100)

This contradicts the claims by Jöreskog and Sörbom (1981) and Werts et al. (1976) that analysis of the correlation matrix violates the distribution assumptions of the likelihood ratio.

The arguments listed above were stated emphatically by their authors and most of these authors appear to rule out the analysis of correlation matrices by MLFA. In practice, however, these authors appear to have made a distinction between the analysis of a correlation matrix for a single sample and simultaneous analysis of several samples to determine equivalence of factor structure. Correlation matrices are analyzed in several confirmatory factor analysis studies by Jöreskog (1969, 1978; Jöreskog & Lawley, 1969; Jöreskog & Sörbom, 1978) and statistical tests are applied in those studies. When a given factor model is tested on two or more samples simultaneously (Jöreskog, 1971; Jöreskog & Sörbom, 1978, 1981; McGaw & Jöreskog, 1971; Werts et al., 1976), the covariance matrix has been analyzed. The critical distinction between these two types of studies may be the importance of homogeneity of variance to decisions regarding the pooling of covariance matrices across samples or the estimation of factor scores for all groups simultaneously. As Cunningham (1978) points out, when covariance matrices are standardized independently for each group, differences in the group variances are not detectable. Analysis of a single sample covariance matrix does not necessarily involve issues of homogeneity of variance across variables. Jöreskog (1979b) states that the sample covariance matrix, S, "may be taken to be a correlation matrix if the model is scale free and if the units of measurement are arbitrary or irrelevant" (p. 446). It is the present author's contention that where a set of tests have been standardized to a common mean and variance and where it is these standardized scores that are of theoretical interest, the units of measurement of the observed variables may be said to be arbitrary and irrelevant. The analysis of the correlation matrix in such a situation would not constitute a violation of the distribution assumptions of the likelihood ratio.

Sample Size

The size of the sample employed in the analysis has implications for both the mathematical calculation of the loss or likelihood functions and the reliability of inferences from the resulting χ^2. In some cases the likelihood function may have no true maximum which satisfies the condition that all unique variance coefficients (ψ_i) are positive (Lawley & Maxwell, 1971). Termed an improper solution, or Heywood case, this situation occurs frequently in the MLFA literature and may be due to the specification of an inappropriate model or sampling error, especially when the sample size is small. Lawley and Maxwell suggest dealing with such occurrences by fixing negative ψ_i to an arbitrary small positive value. Driel (1978) notes that Heywood cases may occur where fewer than three variables load on a given factor and suggests dropping variables which load on underidentified factors. Employment of large samples would presumably be one preventitive measure open to the researcher.

The vaguely-stated third assumption of "relatively large" (Jöreskog & Sörbom, 1981, p. I.39) sample size becomes problematic when one considers that the value of χ^2 for a given model is a direct function of the sample size (χ^2=NF$_0$, see Equation II.9). The result of this relationship is that virtually any model would be considered untenable when very large sample sizes are employed, while accepting a null hypothesis of equality of Σ and S in small samples may be inappropriate (Bentler & Bonett, 1980). Although this is true of any statistic, given Bakan's (1966) assertion that the null hypothesis is almost always false, the fact that support for the factor model is demonstrated by failure to reject the null hypothesis alters the implications of this relationship for concerns about Type I vs. Type II error. Schönemann's (1981) discussion of these implications is presented toward the end of the present section on the statistical assumptions associated with MLFA.

Various solutions have been proposed to account for this sensitivity to sample size. The practice of blaming rejection of one's favored factor model on excessive sample size and ignoring the results (Bejar & Doyle, 1981; McGaw & Jöreskog, 1971) would appear to be the least defensible of these. Jöreskog (1978) has stated that when large sample sizes are employed the researcher should consider the difference between the χ^2 values of successive models rather than the χ^2 values themselves. Jöreskog and Sörbom (1981) recommend a judgemental comparison of the change in χ^2 to the change in degrees of freedom between models. Both the Bejar and Doyle (1981) and the McGaw and Jöreskog (1971) studies reported the acceptance of factor models although less restricted models (e.g., extraction of more factors) resulted in significantly lower χ^2 values for the change in degrees of freedom. Given the theory that the χ^2 distribution of NF_0 holds when sample size is large, this neglect of the implications of even incremental tests of significance appears to be unwarranted. Monte Carlo studies of the behavior of this statistic with various sample and model characteristics are described below and their implications for decisions regarding the number of factors or the factor pattern to accept are discussed.

Browne (1968) compared the accuracy of loadings and number of factors extracted by maximum likelihood, principal components, and principal factor analysis from data generated to fit a four-factor model. Correlation matrices were generated from scores on 12 variables for a sample size of 100; on 16 variables for $N=100$; and on 12 variables for $n=1500$. Browne concluded that likelihood ratio tests for

MLFA tended to underfactor when N=100 and 12 variables were analyzed, as three-factor solutions were accepted for 5 of 20 such samples. Increasing the number of variables to 16 resulted in greater accuracy in the number of factors, while the likelihood ratio indicated four factors for all samples of 1500 with 12 variables. A variant of the likelihood ratio test developed by Jöreskog (1962) tended to overfactor when 12 variables were analyzed, whether samples of 100 or 1500 were used. Jöreskog's test correctly indicated four factors for all samples when 16 variables were analyzed. Browne (1968) also found that increasing either sample size or the ratio of variables to factors increased the accuracy of loadings and decreased the frequency of Heywood cases.

Linn (1968) generated correlation matrices for 20 variables from a seven-factor model and found that MLFA overestimated the number of factors for large samples when the matrix of uniquenesses was not diagonal for the population, i.e., when correlations were present among the error coefficients, or when communalities were very high. This tendency to overfactor was not present when N=100 or the assumptions of the formal factor analytic model were met. Although Linn's design demonstrates a pattern of relationships among sample size, communality size, and the number of factors indicated by MLFA, this demonstration was based on only one factor matrix for each experimental condition. A similar but expanded study, examining these relationships with several hundred simulated correlation matrices, would provide more reliable information on the variables affecting the power and accuracy of MLFA tests on the number of factors.

Geweke and Singleton (1980) tested the power of the likelihood ratio for sample sizes of 10, 30, 150, and 300, employing simulated scores on five variables in models which varied on number and qualitative nature of factors. Contrary to the expectations implicit in the above discussions on sample size by Bentler and Bonett (1980) and Jöreskog and Sörbom (1981), Geweke and Singleton reported that rejection of the null hypothesis (of acceptable model fit, where the data was derived from the model) was rejected with excessive frequency when N=10. Samples of 300 were associated with infrequent rejections of the null hypothesis. Schönemann (1981) suggested that these surprising results were a function of the unrealistically high communalities built into the Geweke and Singleton models. When Schönemann repeated this experiment, lowering the average communality from over .9 to .5 (the average h^2 reported in a review of factor analytic literature), the range of power coefficients dropped from .8-.9+ to .2-.5. These results demonstrate a severe limitation on the generalizability of Geweke and Singleton's conclusions. The effect of high communalities on the small sample results in Geweke and Singleton's study are consistent with the overfactoring which occurred where communalities were high in Linn's (1968) study and with results obtained by Hakstian, Rogers, and Cattell (cited in

Cattell, 1978). However, this does not explain the relative lack of power for large samples in Geweke and Singleton's results. In addition, Schönemann does not consider the problems associated with the underidentification of factors implicit in a two-factor model for five variables. There is a need for more Monte Carlo studies which vary sample size, number of variables, number of factors, and ratios among these variables systematically and test the properties of the likelihood ratio with a large number of sample covariance or correlation matrices in each experimental condition. In the absence of such studies, the power of the statistic for most research applications remains unknown (Bentler & Bonett, 1980).

Faced with incomplete information on the power of the likelihood ratio for small samples, some authors have proposed rules of thumb for determining the sample size needed for tests with a given number of variables. Lawley and Maxwell (1971) suggest that a reasonably large sample size may be defined as $N-p \geq 50$, where N is the sample size and p equals the number of observed variables. Cattell (1978) suggests that, since ML estimates are based on large-sample theory, sample sizes should equal or exceed 80 subjects, with 200 or more as a preferable number. Nunnally (1978) recommends a sample size 10 times as large as the number of observed variables, a more stringent requirement than Lawley and Maxwell's for most social science applications.

Bentler and Bonett (1980) propose a null model against which theoretical models might be tested. The null model states that all variables are mutually independent, i.e. that all covariances equal 0. If a given model does not provide a significantly better fit than the null model, this theoretical model is not informative. When the sample size is large, confidence may be placed in a model for which the χ^2 indicates a poor fit if this model is a significant improvement over the null model. When sample size is small, Bentler and Bonett suggest that concerns regarding sufficiency of power, which may arise when the theoretical model is not rejected, may be allayed if this model is significantly more informative than the null model. This rationale is weakened by the fact that the researcher is not aware of the amount of additional information which is unavailable from the model. Where a given model is not significantly better than the null model, and not significantly poorer than the satiated model, the data is inadequate for model-testing. Such a condition is only likely to occur where sample size is very small.

The difficulties posed for model-evaluation by the sensitivity of associated statistics to sample size led Tucker and Lewis (1973) to derive a reliability index for MLFA. The index is defined as

$$\rho_m = (M_0 - M_m)/(M_0 - 1/n_m) \qquad \text{(II.11)}$$

where M_m is the value of the loss function for a model m divided by the degrees of freedom for that model, M_0 is a similar ratio for the null model, and n_m is the sample size. Tucker and Lewis provide several examples of analyses in which this index supported factor models which had long been accepted by researchers in the field but rejected by the strict application of the χ^2 test of fit.

Jöreskog and Sörbom (1981) have programmed calculation of a similar index into **LISREL** V. The Goodness-of-Fit index is defined as

$$GFI = 1 - tr(\Sigma^{-1}S - I)^2/tr(\Sigma^{-1}S)^2 \qquad (II.12)$$

where I is an identity matrix. These indices are intended to range between 0 and 1, with a large value indicating a good fit for the model tested. However, both ρ_m and GFI may theoretically be negative. Bentler and Bonett (1980) provide an extension of ρ_m which provides a measure of the practical significance of an improvement in fit from a model l over a more restricted model k.

The reliability and goodness-of-fit indices derived by Bentler and Bonett (1980), Jöreskog and Sörbom 1981), and Tucker and Lewis (1973) are promoted by their authors as more reasonable guides to model evaluation decisions than χ^2 tests of significance. These arguments are based on the claims of the various indices to independence from sample size. However, as Jöreskog and Sörbom (1980) warn about the use of GFI, there is no standard against which to judge the value of these coefficients. Their apparent conservatism may indicate that they are particularly inappropriate when achieving sufficient power is a concern, such as in small-sample situations.

Two final notes of caution on the interpretation of MLFA results concerns the logic of inferences involved. Confirmatory factor analysis procedures described by Jöreskog (1969, 1978; Jöreskog & Sörbom, 1978, 1981) include the modification of models which are rejected as untenable on the basis of a significant χ^2. Cliff (1983) argues that a model loses its status as a hypothesis about a set of data once it is adjusted in the light of that same data. The model finally chosen by such a procedure therefore represents an unstable picture of the relationships among the variables. Jöreskog (1969) recognizes the exploratory nature of the sequence of tests on modified models, implying at least partial agreement with Cliff's argument. Even when a modified model results in significant improvement, this modification is not necessarily generalizable to the population. Cliff recommends cross-validation of results on a new sample to test the reliability of the final model.

Given the indeterminacy of factor scores discussed in the beginning of this discussion on factor matching, a factor model which explains the data well is only one of several models which may be adequate. Cliff (1983) notes that when a model is not rejected, it simply means that that model is tenable, not that it is true. Schönemann (1981) concurs with Cliff in warning that cognitive theorists abuse factor analysis by presenting factor patterns as proof of the occurrence of mental processes described in their theories. Schönemann also argues that MLFA methods of confirmatory factor analysis control probabilities of Type I error (that of incorrectly rejecting an accurate model), but that insufficient attention is directed to avoiding Type II error (failing to reject a poor model). The onus should be on the apologist for a theory to support the factor model, rather than simply failing to disconfirm it. This kind of evidence requires methods other than factor analysis. Such evidence will be discussed in Chapter V of this report.

Having considered the hypotheses which may be tested by MLFA and some of the constraints upon those tests, the focus of this chapter shifts to the strategy of sequencing the tests of various hypotheses about a data set.

Sequence of Analyses

When a researcher wishes to examine the fit of a factor model to a collection of samples, the evaluation of the model's validity may be divided by testing two major hypotheses. The first hypothesis is that the collection of samples have equivalent observed score covariance matrices which may therefore be explained by identical factor models. The second hypothesis is that the theoretical factor model of interest explains this common covariance matrix. Testing of each of these major hypotheses includes a series of tests on various sequential hypotheses, which are described below.

Simultaneous Factor Analysis

The sequence for simultaneous factor analysis of several populations (SIFASP), as developed by Jöreskog (1971), begins with the test of homogeneity of covariance matrices across populations. This test may be expressed as

$$H_{0-1}: \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k \qquad \text{(II.13)}$$

for K populations. If this null hypothesis is tenable, the covariance matrices may be pooled and further analysis conducted on the pooled matrix. If H_{0-1} is rejected, the similarities and differences in factor structure may be investigated in the manner described below.

If the covariance matrices cannot be considered equivalent, the number of factors is tested simultaneously across groups. This test is expressed as

$$H_{0-2}: M_1 = M_2 = \ldots = M_k = M \qquad \text{(II.14)}$$

where M is the number of factors expected to be significant. For example, if theory and/or previous studies had suggested that three factors were sufficent to explain a given set of variables, the sufficiency of three factors for all populations would be tested without constraining the size of the loadings to be equal across groups (Jöreskog, 1971).

If the hypothesis of an equal number of factors across groups is tenable, the equivalence of factor patterns is tested. Expressed as

$$H_{0-3}: \Lambda_1 = \Lambda_2 = \ldots = \Lambda_k \qquad \text{(II.15)}$$

this hypothesis may require the size of corresponding loadings to be equal across group factor matrices or may simply require that the same set of loadings be identified as salient. In the former case, the researcher specifies that all salient loadings are equal and that all nonsalient loadings equal 0. In the latter case, the researcher specifies that all nonsalient loadings are 0, but allows the salient loadings to be estimated by the maximum likelihood procedure. If the χ^2 obtained by testing H_{0-3} against H_{0-2} is not significantly larger than expected by chance, the hypothesis of homogenous covariance matrices need not be rejected. Otherwise, the factor patterns for the various groups must be examined separately for discrepancies from the hypothesized pattern.

If the factor patterns may be considered invariant across groups, the subsequent hypothesis tested is that both factor patterns and uniqueness coefficients are invariant across groups. This hypothesis is expressed as

$$H_{0-4}: \Lambda_1 = \Lambda_2 = \ldots = \Lambda_k; \quad \Psi_1 = \Psi_2 = \ldots = \Psi_k \tag{II.16}$$

and is tested against the χ^2 obtained for H_{0-3}. The test of invariance of factor intercorrelation matrices is expressed as

$$H_{0-5}: \Lambda_1 = \Lambda_2 = \ldots = \Lambda_k; \quad \Phi_1 = \Phi_2 = \ldots = \Phi_k; \quad \Psi_1 = \Psi_2 = \ldots = \Psi_k. \tag{II.17}$$

This hypothesis assumes that H_{0-4} has not been rejected and is tested against the χ^2 obtained for that hypothesis.

If, at any point in the above procedure, the null hypothesis of interest is rejected, the researcher would explore the differences in models across groups by altering the specification of parameters thought to be poorly specified by the hypothesized factor model. The detection of such parameters is described in the section on confirmatory MLFA which follows. This section is primarily concerned with testing the fit of a hypothesized factor model to the covariance or correlation matrix of a single sample or the pooled matrix for a collection of samples.

Confirmatory Analysis with a Single Sample

The procedures and conditions for confirmatory factor analysis have been set out by Jöreskog (1969). More recent guides (Kroonenberg & Lewis, 1982; Lomax, 1982) extend and refine Jöreskog's suggestions but retain his basic logic of inquiry. After determining that the number of factors in the theoretical model are sufficient to explain the sample covariance matrix, the model is tested and, if untenable, modified until an acceptable fit to the data is obtained. Before describing these steps in more detail, two distinctions between types of factor models must be noted.

A unique solution is defined as one in which all linear transformations which leave the fixed parameters unchanged also leave the free parameters unchanged. If a given solution is nonunique, standard errors may not be calculated for estimates of free parameters (Jöreskog & Sörbom, 1978, 1981). A minimum condition for uniqueness is the presence of

M^2 fixed parameters in the common factor space, where M is the number of common factors in the model (Jöreskog, 1969).

An unrestricted factor solution does not restrict the common factor space, i.e, does not restrict $\Lambda\Phi\Lambda'$. A restricted solution places restrictions on the factor space and cannot be obtained by rotation of an unrestricted solution for the same data (Jöreskog, 1969). A restricted solution has more than M^2 restricted parameters, yet both restricted and unrestricted solutions may be unique or nonunique. The choice of parameters to be fixed in unrestricted models is explained in detail in Chapter III.D. The parameters fixed in restricted models are chosen on the basis of the factor theory being tested.

The sufficiency of M factors may be examined by testing an unrestricted unique solution for M factors. If the hypothesis that M factors is sufficient is considered untenable, the hypothesized factor pattern for M factors will likely be untenable. If M factors may be considered sufficient, the hypothesized factor pattern may be tested.

The hypothesized factor pattern is tested by freeing all loadings in Λ which are expected to be salient for estimation by the maximum likelihood procedure and fixing all other loadings to 0. If the model is orthogonal, the off-diagonal elements of Φ are fixed at 0; if oblique, they are free to be estimated. The χ^2 obtained for the model provides an index of it's fit to the data. The goodness of fit may also be evaluated by examining the standard errors of the parameter estimates. If the model is rejected, the analysis becomes exploratory, rather than confirmatory, as the researcher modifies the model to arrive at a solution which provides a satisfactory fit to the data. A number of clues for identifying the parameters to be modified are available from the output of MLFA programs such as LISREL and these indices are discussed below.

Parameters which were free and significant in the unrestricted solution but fixed to 0 in the restricted solution may be freed to examine their significance in the restricted solution (Jöreskog, 1969). These parameters may be identified by their confidence intervals in the unrestricted solution.

The largest elements of the matrix of residual variances and covariances, i.e., the matrix of differences between corresponding elements of S and Σ, may suggest relationships among the variables which have not been accounted for by the model tested (Jöreskog & Sörbom, 1981). Sörbom (1975) argues that the residuals may be inappropriate guides for decisions on maximum likelihood estimators and are more appropriate for models with least squares estimators. Lomax (1982) suggests that observed variables with a large number of large residuals should be examined for low reliability and discarded if such measurement problems seem apparent.

Partial derivatives are calculated for all fixed parameters in a model. The size of this derivative is an index of the amount of distortion in the model attributable to fixing a given parameter to a given value. Restrictions on parameters associated with large derivatives may be relaxed and the incremental improvement to the model's fit to the data noted (Jöreskog & Sörbom, 1978; Sörbom, 1975). Parameters are freed one at a time, as freeing one parameter may radically shift the ranking of partial derivatives among the remaining fixed parameters (Jöreskog & Sörbom, 1981; Lomax, 1982; Sörbom, 1975). **LISREL** V (Jöreskog & Sörbom, 1981) calculates modification indices, based on the partial derivatives, such that the parameter with the largest modification index will result in the largest decrease in χ^2 if freed.

Parameter estimates which are nonsignificant may be fixed to 0 and the model retested. Jöreskog (1969, 1978) suggests that parameters with 95% confidence intervals which include 0 should be considered nonsignificant. Kroonenberg and Lewis (1982) argue that a 99% confidence interval should be used since these intervals are considered individually rather than simultaneously. **LISREL** V calculates t-values for each parameter by dividing the estimate by its standard error. If estimates are normally distributed (Lawley & Maxwell, 1971), a parameter with a t-value exceeding 2.0 will have a 95% confidence interval which excludes 0 (Jöreskog, 1978; Jöreskog & Sörbom, 1981; Kroonenberg & Lewis, 1982).

The indicators discussed above all provide clues to the adjustment of a factor model at one step in the analysis. They do not provide an overall strategy for the analysis, which might guide decisions on initiating and terminating model modification. Bentler and Bonett (1980) describe a hierarchical set of models to be tested, in which each restricted model is a special case of a less restricted model. In other words, a restricted model, f_t, may be tested against a less restricted model, f_s, where the vector of parameter estimates Θ_t is a subset of the vector of estimates Θ_s. The satiated model defines Σ as any positive definite matrix and serves as the H_1 against which the overall fit of any model is tested. The null model, which defines Σ as a diagonal matrix, is the most restricted model in the hierarchy. All other models fall between these two, being nested in the satiated model and having the null model as a subset of its estimated parameters. Models of the factor structure are subsets of the unrestricted models for testing the sufficiency of a given number of factors. The unrestricted model for testing the sufficiency of two factors should be a special case of the model for testing the sufficiency of three factors. The degree to which this sequential nesting of models, or parameter vectors, is practised determines the extent to which the researcher may apply tests across various levels of models.

There appears to be a consensus among the authors cited thus far that only parameter modifications which fit some theory should be tested (Jöreskog, 1969; Lomax, 1982; Kroonenberg & Lewis, 1982; Sörbom, 1975). Although this practice may lend conceptual order to the model-fitting sequence, it is the present author's contention that it may result in the loss of information which might shed further doubt on the original model or shed further light on possible explanations for that model's poor fit. This result is of particular concern given the tendency of some researchers to blame large χ^2s for their model on large sample sizes and claim support for the model on theoretical grounds (Bejar & Doyle, 1981; Cunningham, 1981; McGaw & Jöreskog, 1971).

Kelderman, Mellenbergh, and Elshout (1981) recommend comparing the model of interest to some alternate theoretical model. However, if the alternate model is not nested within the model of interest, there is no basis for judging the difference in χ^2 values obtained for the two models.

A more rigorous procedure might involve making an a priori decision on exactly what modifications are admissable without invalidating the theory being tested. For example, the Arithmetic subtest of the WISC-R is interpreted primarily as a a Freedom From Distractibility measure, yet simultaneous loadings on the Verbal Comprehension factor would not be inconsistent with Kaufman's (1975) results for the standardization sample or his guidelines for interpreting the FD subtests in the context of Verbal and

Performance Scale scores (Kaufman, 1979). If the theoretical model(s) do not fit the data, i.e. the χ^2 is significantly large or critical loadings are nonsignificant, the researcher could state that the model did not sufficiently explain the relationships among the variables for that population. Whether subsequent modifications are based on some alternate theory about the variables and the population or on "blind empiricism", conclusions based on the results of the modifications must be considered unreliable until cross-validated (Cliff, 1983).

As demonstrated by the studies cited above, the strategy of analysis by MLFA may vary according to the nature of the questions asked about the model and the assumptions associated with those questions. The power of MLFA to provide reliable and valid evidence for questions concerning factor structure is summarized below and compared to other confirmatory factor analytic methods.

Comparison to Other Confirmatory Methods

Maximum likelihood methods of factor analysis provide the researcher with a χ^2 statistic which indicates the goodness of fit of the covariance matrix for a sample to the covariance matrix implied by the factor model being tested. Confirmatory MLFA is similar to Procrustes rotations to maximal congruence of factor matrices in that a target factor matrix is employed. MLFA differs from the latter method in that exact loadings for the target matrix need not be specified (Nunnally, 1978). Although initial parameter estimates must be entered for many MLFA programs, the test of the factor model may be restricted to the pattern of salient vs. nonsalient loadings.

Maximum likelihood methods of confirmatory analysis are similar to the multiple group method in that a direct factor solution is applied to the correlation matrix (Nunnally, 1978). Nunnally suggests that multiple group methods are preferable for ease of calculation and conceptualization. However, multiple group factors which include few variables will have loadings which are inflated by the large weighting of each member test on the factor score.

One advantage of confirmatory MLFA is its facility for allowing inferences about error coefficients and the correlations among factors. Alwin and Jackson (1981) argue that tests for confirmatory factor analysis or factor invariance across groups which only account for the factor pattern confound sources of variance among factor models. The only other confirmatory method which appears to deal with factor intercorrelations directly is the confactor method of Kaiser et al. (1971).

The confidence intervals available from MLFA procedures provide further evidence for the test of validity of a factor model for a given group. Where the χ^2 for a model is not significant but several theoretically important parameters are also nonsignificant, modification to the model may be indicated.

One important weakness of MLFA is its apparent sensitiviy to sample size and to violations of the assumption of multivariate normal score distributions. Although consensus is lacking on the optimum sample size to employ for a given number of variables, most of the Monte Carlo studies have suggested that increasing sample size leads to increased accuracy in the values of loadings and decisions on the number of factors present. The effect of sample size appears to be mediated by the size of the communalities and adherence of the population data to the formal factor model. Several procedures for determining the fit of a model, or the significance of improvement from changes to a model, have been proposed by Bentler and Bonett (1980) and by Jöreskog and Sörbom (1981) and reviewed in earlier sections of this report.

Comparisons to other factor extraction and rotation procedures tend to favor MLFA over other methods, particularly for confirmatory factor analysis. Maximum likelihood estimates of loadings were judged to be of equivalent or superior accuracy in comparison with those of principal factor and component analysis (Browne, 1968; Linn, 1968). Cattell (1978) reports that MLFA results tend to agree with those of principal factor analysis, but suggests that MLFA and unweighted least-squares analysis (which is not discussed in this report) provide the most refined factoring results. Nunnally (1978) contends that MLFA is not as subject to capitalization on chance as Procrustes rotation.

The statistical properties summarized above, plus the diagnostic information available from the partial derivatives of fixed parameters, places MLFA at an advantage over the other confirmatory methods and indices discussed in this chapter. Consequently, this method was chosen to test the validity of the factor models proposed in Chapter II.B as interpretive guides for the WISC-R scores of Inuit children. The main hypotheses are outlined below. The specification of parameters for the sequences of nested hypotheses is described in Chapters III and IV.

E. The Present Study

As stated in Chapter I, the primary purpose of the present study is to determine whether three factor models for clinical interpretation of the WISC-R may be applied to the test scores of children, aged 7-0 to 14-11 years, from the Districts of Keewatin and Kitikmeot, Northwest Territories. The factor models to be tested are those defined by the organization of subtests into the Verbal and Performance Scales by Wechsler (1974); the Verbal Comprehension, Perceptual Organization, and Freedom from Distractibility factors, by Kaufman (1975); and the categories of Spatial, Conceptualization, Sequential, and Acquired Knowledge, by Bannatyne (1974). The validity of the factor models will be tested by applying maximum likelihood factor analysis to the covariance matrices of the NWT samples. The χ^2 goodness of fit statistic provided for each model, plus the significance of all theoretically important parameters, will be the basis of decisions regarding that model's validity for the age group in question. Significant departure from the factor structure implied by a clinical model will be interpreted as evidence against the use of that model for the interpretation of WISC-R subtest scores of children from the two NWT districts.

A secondary purpose of the present study is to arrive at factor models which explain the subtest correlations for those age groups for whom one or more of the clinical models is inappropriate. The diagnostic guides to sources of model distortion, described in Chapter II.D will be employed to modify inappropriate models until a satisfactory fit to the data is achieved. The secondary status of the exploratory analysis is a function of the lower reliability of the conclusions which may be drawn from the results, pending cross-validation with other samples. Interpretations for such results will be offered as hypotheses rather than conclusions.

The specific hypotheses to be tested are described below. The manner in which the parameters in Λ, Φ, and Ψ will be specified to allow these tests will be described in Chapter III.D.

1. The scores on each subtests for each age group will be normally distributed with a mean of 10.0 and a standard deviation of 3.0, i.e., N(10,3).

The Kolmogorov-Smirnov one-sample test statistic (Massey, 1951; Smirnov, 1948) will be calculated to determine the fit of each score distribution to normality. As noted in Chapter II.D, one assumption of the likelihood ratio and associated statistics is that scores on all variables have a multinormal distribution. Departures from normality will limit confidence in the factor analysis results pertaining to those subtests.

The following hypotheses regarding equivalence of covariance matrices are tested with the simultaneous factor analysis (SIFASP) procedures developed by Jöreskog (1971; Jöreskog & Sörbom, 1978, 1981) and described in Chapter II.D of this report.

2. The covariance and correlation matrices for the eight age groups are equivalent. This hypothesis may be expressed as

   $H_0: \Sigma_7 = \Sigma_8 = \Sigma_9 = \Sigma_{10} = \Sigma_{11} = \Sigma_{12} = \Sigma_{13} = \Sigma_{14}$

   where the subscripts 7-14 refer to the age group represented by the respective covariance matrix.

   If the hypothesis is tenable, the covariance matrices for the eight age groups will be pooled and transformed to a correlation matrix. All of the clinical models would then be tested on this pooled matrix as well as on the matrices for separate age groups or two-year age pools.

3. Successive pairs of age group covariance matrices are equivalent. This set of hypotheses may be expressed as

$H_0: \Sigma_7 = \Sigma_8$

$H_0: \Sigma_9 = \Sigma_{10}$

$H_0: \Sigma_{11} = \Sigma_{12}$

$H_0: \Sigma_{13} = \Sigma_{14}$.

Although this set of hypotheses is implied by the hypothesis of equivalence of all eight covariance matrices, the latter hypothesis may be false without negating the pair-wise hypotheses.

If these hypotheses are tenable, the covariance matrices within each of the pairs tested would be pooled and further confirmatory factor analysis would be conducted on these four age-group pools. Pooling these two-year age groups increases the power of the statistic to identify a poor fit to a given model, while still allowing examination of developmental trends in the factor structures that emerge. If any of the above pair-wise hypotheses are rejected, the data of the age groups concerned will be analyzed separately.

The sequence of hypotheses described below is based on the procedures for confirmatory factor analysis developed by Jöreskog (1969, 1979b; Jöreskog & Sörbom, 1978) and described in Chapter II.D of this report. These hypotheses will be tested for the pooled correlation matrix for the total sample if the hypothesis under 2 above is not rejected. They will be tested for the four age pool matrices if the hypotheses under 3 above is not rejected.

4. A single factor is sufficient to explain the correlation

matrix.

Rejection of this hypothesis will not be interpreted as evidence that a general factor does not affect the subtest score variance, but as evidence that two or more primary factors are required to explain the correlations. If those factors are oblique, the nature of the general factor(s) could be explored through hierarchical factor analysis. Although such analysis of U.S. normative data for the WISC-R has been conducted, as described in Chapter II.B, the results of these studies have not been incorporated into the most prevalent interpretive models for the test. Therefore, the present study will not include examination of second-order factors underlying the NWT data.

5. Two factors are sufficient to explain the correlation matrix.
6. Two factors do not improve upon the fit of a single general factor.

The two-factor solution proposed in the two tests above is an unrestricted unique solution as defined by Jöreskog (1969, 1979a, 1979b) and described in Chapter II.D of this report.

7. The Verbal and Performance Scales defined by Wechsler (1974) explain the subtest correlations as a restricted two-factor model.
   a. The Verbal and Performance factors are orthogonal.
   b. The Verbal and Performance factors are oblique.

If both of these hypotheses are rejected although two factors are sufficient, the various clues to model distortion described in Chapter II.D will be examined and the model will be sequentially modified until an acceptable fit to the data is achieved.

8. Three factors are sufficient to explain the subtest correlations.
9. The three factor solution is not a significant improvement upon the fit of a two-factor solution.

The two hypotheses just stated regarding the sufficiency of three factors refer to an unrestricted solution. This solution will be compared to an unrestricted two-factor solution, as the final restricted two-factor solution may not be a special case of the final three-factor solution. If a two-factor solution is adequate, but three factors significantly improve the fit to the data, the argument for accepting the final two-factor model will be weakened. This decision will depend in part on the relative interpretability of the final two- and three-factor solutions.

10. Kaufman's VC, PO, and FD factors explain the subtest intercorrelations.
    a. The three factors are orthogonal.
    b. The three factors are oblique.

As with the restricted two-factor model, the test of Kaufman's model involves a restricted three-factor solution. If both the orthogonal and oblique models are rejected, the model will be further modified until a satisfactory fit to the data has been achieved, presuming three factors are sufficient.

11. Four factors are sufficient to explain the correlation matrix.
12. Four factors do not improve upon the fit of three factors.

The last two hypotheses above refer to an unrestricted four-factor solution, which will be compared to an unrestricted three-factor solution.

13. Bannatyne's (1974) Spatial, Conceptualization, Sequencing, and Acquired Knowledge categories define a restricted four-factor model which explains the subtest intercorrelations.
    a. The four factors are orthogonal.
    b. The four factors are oblique

    If both the orthogonal and oblique models are rejected, the model will be modified according to the indices described in Chapter II.D.

Should the hypothesis that all eight covariance matrices for the respective age groups are equivalent be tenable, one final hypothesis will be tested for each of the four age pools.

14. The final factor solution(s) derived for the total sample will explain the pooled correlation matrix of

each two-year sample. This hypothesis implies not only a nonsignificant χ^2 for the model(s), but that all free parameters which were significant for the total group data will be significant for the two-year age pools. The above test is not equivalent to performing a cross-validation of the results of each sample to a new sample. However, this test does provide some additional information regarding the stability of the total sample results across the various age groups.

III. Methodology

A. Subjects

Description of Sample

The sample consisted of 366 children from 13 schools in the Districts of Keewatin and Kitikmeot in the Northwest Territories, Canada. Samples were drawn independently for each year in the age range 7-0 years to 14-11 years. Sample sizes for the various age levels ranged from 34 to 53 children. The number of children representing each school within each age group is presented in Table 1. The method of selection and descriptions of the communities involved are provided below.

The populations of interest to this study are defined as all children on the school registry of the NWT Department of Education whose age fell within one of the specified ranges at the time of testing. Seventy-two children (36 males, 36 females) were drawn from this list at each age level, using a computer-generated table of random numbers. The first 50 of these were scheduled to be tested; the final 22 children were selected as replacements for any original subjects who were unable to participate.

Description of Communities

The schools attended by these children were located in 13 communities spread along the west shore of Hudson Bay,

Table 1

Final Sample Distribution By Village, Sex, and Age[1]

		Age Groups								
Village	Sex	7	8	9	10	11	12	13	14	TOTAL
Baker Lake	M	3	3	2	5	3	2	1	2	21
	F	2	3	3	2	2	0	2	3	17
Chesterfield Inlet	M	0	1	2	0	2	0	1	1	7
	F	1	0	0	1	3	1	2	0	8
Coral Harbor	M	1	1	6	2	0	0	3	1	14
	F	0	1	2	3	1	4	3	1	15
Eskimo Point	M	1	0	1	4	4	3	4	4	21
	F	2	3	4	2	4	4	5	3	27
Rankin Inlet	M	0	3	2	3	3	2	3	4	20
	F	3	2	3	3	4	3	1	4	23
Repulse Bay	M	3	0	2	0	2	3	1	1	12
	F	1	0	0	1	1	1	1	0	5
Whale Cove	M	0	1	2	0	1	3	1	1	9
	F	0	0	0	0	0	0	0	1	1
Cambridge Bay	M	1	2	2	2	5	1	3	2	18
	F	1	2	3	3	2	2	2	2	17
Coppermine	M	3	2	1	4	1	3	5	2	21
	F	2	3	3	2	1	3	1	2	17
Gjoa Haven	M	1	2	2	2	2	4	3	2	18
	F	2	3	3	2	4	5	3	1	23
Holman Island	M	0	2	0	2	0	2	2	0	8
	F	1	1	0	0	1	1	3	2	9
Pelly Bay	M	1	1	0	2	3	1	2	0	10
	F	2	0	0	1	3	2	1	1	10
Spence Bay	M	3	1	2	0	0	0	0	0	6
	F	0	0	3	3	0	0	0	3	9
Total Sample	M	17	19	24	26	26	24	29	20	185
	F	17	18	24	23	26	26	24	23	181
Total N = 366										

[1] Adapted from Mulcahy & Watters, 1982, p. 6.

the Arctic coastline, and Victoria Island. All of the communities are located north of the tree line on tundra or sandy terrain. A brief description of each village is provided to give the reader a clearer picture of the setting in which the children live. Demographic information was obtained from an annual survey of government and community sources of northern demographic information (Devine & Wood, 1981). Information on language instruction in the schools was provided by the Department of Education of NWT (Watters, Note 1). Although population figures are presented for 1979, information regarding ethnic representation in the communities was only available for 1978.

District of Kitikmeot

Holman Island

This settlement is located on the western edge of Victoria Island at 70°44′N, 117°44′W, 575 miles north of the territorial capital of Yellowknife. Its 1979 population was 336, 88% of whom were Inuit. The principal languages spoken are Inuktitut and English. The main economic activities are print-making, trapping, hunting, sealing, fishing, and oil and gas exploration. The town receives scheduled air service from Yellowknife, postal and telephone service, and CBC radio and television. A barge operating from Hay River, NWT visits once per year and bulk shipments are often received in this way.

School enrollment in Grades K-9 was 95 for the 1980-1981 term, and the school employed four teachers and two classroom assistants. English is the language of instruction. Inuktitut is taught as a second language for approximately 30 minutes a day to students in all grades.

Coppermine

Located on the mainland Arctic coast, 350 miles north of Yellowknife at 67°50′N, 115°05′W, this hamlet's main economic activities are handicrafts and carving, trapping, hunting, fishing, and oil and gas exploration. Its 1979 population of 766 was approximately 92% Inuit. Inuktitut and English are the major languages spoken. Coppermine receives scheduled air service, annual barge service, telephone, radio, and television communication.

Grades K-9 are taught at the local school, where the 1980-1981 enrollment was 243 and the staff included 12 teachers and four classroom assistants. English is the language of instruction. The classroom assistants deliver an Inuktitut program to students.

Cambridge Bay

This settlement lies on the south-eastern edge of Victoria Island, 538 miles north-east of Yellowknife at 69°07′N, 105°03′W. Approximately 77% of its 1979 population of 864 was Inuit, with 1% Dene. Inuktitut and English are the principal languages. Cambridge Bay serves as a regional government center and a base for a Distant Early Warning

(DBW) Line station. Other major economic activity includes commercial fishing, trapping, transportation and communications services. Scheduled air service is provided via Yellowknife as well as local charter service. Barge service is provided according to demand.

Grades K-9 are available and 247 students were enrolled in the 1980-1981 term. Staff included 13 teachers and 5 classroom assistants. English is the language of instruction. Elementary students receive 30 minutes per day of instruction in Inuktitut as a second language.

Gjoa Haven

This hamlet is located on the coast of King William Island, just off the mainland Arctic coast. At 68°38′N, 95°53′W, it lies 660 miles northeast of Yellowknife. Its 1979 population was 493. Approximately 93% of these were Inuit, and both Inuktitut and English are spoken. The main economic activities are hunting, trapping, fishing, carving and handicrafts. Scheduled air service is provided via Cambridge Bay and annual barge service is available. Telephone, radio, and television service are provided.

Grades K-8 are taught at the local school, where the 1980-1981 enrollment was 187. Ten teachers and 4 classroom assistants were on staff. Inuktitut is the language of instruction for Kindergarten children. English is introduced as an instructional language in Grade 1 and the proportion of instruction delivered in Inuktitut is decreased as the children advance through the various grade divisions.

Spence Bay

Located 765 miles north-east of Yellowknife and 288 miles east of Cambridge Bay, at 69°32′N, 93°32′W, Spence Bay had a 1979 population of 470. Approximately 93% of these were Inuit and both Inuktitut and English are spoken. Carving and other handicrafts, commercial fishing, trapping, and hunting comprise the main economic activities. Scheduled air service is available via Cambridge Bay, and mail, telephone, local and CBC radio and CBC television are available.

Grades K-7 are available and the 1980-1981 enrollment was 164. Eight teachers and four assistants are on staff. Kindergarten and Grade 1 children receive instruction in Inuktitut. Children in Grades 2 and 3 have the aid of Inuit classroom assistants, while older children receive 30 minutes of Inuktitut instruction per day.

Pelly Bay

This hamlet is located on the Arctic coast mainland at 68°32′N, 89°48′W, 815 miles north-east of Yellowknife. Its 1979 population was 281, approximately 93% of which were Inuit. Inuktitut and English are the principal languages. Commercial fishing, hunting, carving and handicrafts are the main economic activities in Pelly Bay. Scheduled air service is available via Cambridge Bay, but the hamlet is inaccessible to barge traffic. Telephone, television, and local and CBC radio service is available.

Grades K-6 are taught locally to 98 children (1979 figures) by five teachers and two classroom assisstants. Inuktitut is the language of instruction for Kindergarten children and older children receive one hour of instruction in Inuktitut per day.

District of Keewatin

Repulse Bay

The hamlet of Repulse Bay lies north of the west coast of Hudson Bay, at 66°32'N, 86°15'W. This is 885 miles north-east of Yellowknife. Its 1979 population was 328 and approximately 92% of its residents are Inuit. Its main economic activities are marine mammal harvesting, hunting, fishing, trapping, carving and other handicrafts. Scheduled air service is available via the regional government center of Rankin Inlet and barge service is provided from Montreal. Mail and telephone service are available and a radio/television receiver was scheduled for service in 1981 or 1982.

Grades K-9 are taught locally. The 1979 enrollment was 126 with five teachers and four assistants on staff. Inuktitut is the language of instruction for Kindergarten and Grade 1 children. Children in Grades 2 and 3 receive 50% of their instruction in Inuktitut and older children receive short daily periods of Inuktitut instruction.

Coral Harbor

Located on the south-west coast of Southampton Island in Hudson Bay, Coral Harbor is 975 miles northeast of Yellowknife and 450 miles west of Frobisher Bay, Baffin Island. Its 1979 population of 414 was approximately 87% Inuit. Inuktitut and English are the primary languages. The major economic activities are marine mammal harvesting, hunting, trapping, transportation and communications. The hamlet was an air force base during World War II and the federal Ministry of Transport still operates a large airfield which employs several local Inuit. Scheduled air service is available via Rankin Inlet and a barge service is operated out of Churchill, Manitoba.

Grades K-9 were taught to 148 children in the 1980-1981 term. The staff included eight teachers and seven full- or part-time classroom assistants. Coral Harbor follows the pattern of Inuktitut instruction provision described for Repulse Bay.

Baker Lake

The only inland Inuit community in NWT, Baker Lake is located at 64°18′N, 96°03′W, 160 miles north-west of Rankin Inlet and 588 miles north-east of Yellowknife. Its 1979 population of 1,017 was 86% Inuit and 1% Dene. Inuktitut and English are the primary languages. The major economic activities are art, hunting, fishing, trapping, tourism, and uranium exploration. Scheduled air service and barge service is available via Churchill. Telephone, local and CBC radio, and television are available

The 1979 enrollment for Grades K-8 was 277, with 15 teachers and 9 full- or part-time classroom assistants on staff. Inuktitut is the language of instruction for Grades K-3, while Grade 4 children receive 50% of their instruction in Inuktitut. Older children receive 30 minutes per day of Inuktitut instruction.

Chesterfield Inlet

This hamlet is located on the west shore of Hudson Bay at 63°21′N, 90°02′, 63 miles north of Rankin Inlet and 713 miles east of Yellowknife. Its 1979 population of 281 was approximately 91% Inuit and both Inuktitut and English are spoken. Major economic activities include hunting, commercial fishing, trapping, and carving. Scheduled air service is available via Rankin Inlet and barge service is provided via Churchill. Residents have access to telephone, local and CBC radio, and television.

There were 85 students in Grades K-9 in 1980-1981 school staff included four teachers and two assistants. Chidren in Grades K-2 receive 50% of their instruction in Inuktitut, while older children receive 45 minutes per day of instruction in this language.

Rankin Inlet

Rankin Inlet is located 295 miles north of Churchill and 715 miles east of Yellowknife at 62°49′N, 92°05′W. Approximately 72% of its 1979 population of 956 was Inuit.

Rankin was a centre for nickel mining until 1962 and operated on a wage economy until the mine closed that year. Its current role as regional headquarters for the territorial government has renewed economic activity in the last decade. Currently, these activities include government administration and services, commercial fishing, transportation, communications, carving and other handicrafts, and trapping. Scheduled air service is available from two air lines via Churchill and Yellowknife, charter air service and barge service are available, and regular overland travel to Eskimo Point is provided in winter. Radio, television, and telephone services are available and radio and television programming are locally produced.

Enrolment in Grades K-9 was 322 in 1980-1981. Thirteen teachers and eight classroom assistants were on staff. English is the language of instruction for at least 50% of the day for the lower elementary grades, while senior classes may receive up to 30 minutes per day of Inuktitut instruction.

Whale Cove

Whale Cove lies 50 miles south of Rankin Inlet and 708 miles east of Yellowknife at 62°11′N, 92°36′W. Its 1979 population of 203 was approximately 92% Inuit and both Inuktitut and English are spoken in the community. Hunting, fishing, and trapping are the main economic activities. Scheduled air service is available via Rankin Inlet and barge service via Churchill. Telephone and radio service are available.

The school has three teachers and two part-time assistants on staff and offered Grades K-9 to 66 students in 1980-1981. Inuktitut is the language of instruction for Kindergarten children and for 50% of the instruction for Grades 1-3. Older children receive short daily periods of Inuktitut instruction.

Eskimo Point

Eskimo Point lies 150 miles south of Rankin Inlet and 675 miles east of Yellowknife, at 61°07'N, 94°03'W on the west coast of Hudson Bay. Its 1979 population of 980 was approximately 94% Inuit and both Inuktitut and English are spoken. The Inuit Cultural Institute is active in educational activities to preserve traditional Inuit culture and have been involved in developing standard written orthographies for Inuktitut. Many of the residents were miners in Rankin Inlet in the 1950s and a gold mine at Cullaton Lake, which is 90 miles inland from Eskimo Point, is expected to provide employment for residents. The main economic activities at present are trapping, hunting, fishing, handicrafts, the Inuit Cultural Institute, and mineral exploration. Devine and Wood (1981) note that Eskimo Point has the largest number of independent local entrepreneurs in the Keewatin. Scheduled air and barge service are available via Churchill. Telephone, local and CBC radio and television service are available. The Cultural

Institute publishes occasional printed materials.

Grades K-9 were offered to 352 students by 15 teachers and seven full- or part-time assistants in 1980-1981. Inuktitut is the language of instruction for children in Kindergarten and Grade 1. Children in Grades 2-3 receive 50% of their instruction in Inuktitut, while older children are given short daily periods of instruction in this language.

A Cautionary Note

While some generalities may be inferred from the above descriptions, the communities differ in such dimensions as length of sustained contact with European or southern Canadian society, committment to preservation of Inuit culture and means of livlihood, and the degree of influence exerted by southern industry and institutions. As mentioned in Chapter II, migration to larger centres such as Rankin Inlet and Eskimo Point has almost doubled the populations of these communities within the last decade, while smaller centres have not increased in size. Demographic trends in NWT are difficult to interpret from such annual statistics as many families leave the settlements in summer to hunt in the interior of NWT. Ethnic representation in the communities would be expected to be sensitive to such seasonal variation. Personal communication with residents of NWT indicated that many residents supplement hunting or fishing with social assistance in the off-season. Many of the entrepreneurships are diversified, e.g. a single family operating taxi and mechanical repair services out of their hardware store. Northern lifestyles and occupations are not easily categorized. The above descriptions are intended to give the reader a quick, if superficial, sketch of the setting in which the testing was conducted.

B. WISC-R

Reliability

The format and interpretation of the WISC-R were discussed in detail in Chapter II.B. The present section is largely restricted to summarizing the reliability statistics presented in Mulcahy and Watters (1982). These results are briefly compared to the reliability indicators presented for the U.S. standardization sample (Wechsler, 1974).

Reliability coefficients were calculated by the method of splitting each subtest into odd- and even-numbered items, correlating the total scores of these two new tests, applying the Spearman-Brown formula to correct for the reduced length of the two half-tests, and presenting the corrected correlation coefficient as a measure of internal consistency. This coefficient was calculated for each age year for all subtests except Digit Span and Coding. The former subtest is administered as two subtests, while the latter is a speeded test. Table 2 presents the range and median reliability coefficient for each subtest in the NWT sample. The corresponding coefficents are presented from the U.S. sample within the age range of 7 to 14 years.

Table 2

Split-half Reliability Coefficients:
NWT[1] and U.S.[2] Norming Samples
Ranges and Median Across Ages 7 Years to 14 Years

	NWT		U.S.	
Subtest	Range	Median	Range	Median
Information	.53-.92	.85	.80-.90	.85
Similarities	.87-.92	.89	.79-.84	.80
Arithmetic	.67-.85	.79	.69-.81	.78
Vocabulary	.83-.90	.89	.70-.91	.86
Comprehension	.75-.86	.82	.70-.87	.80
Picture Completion	.75-.93	.84	.68-.85	.77
Picture Arrangement	.56-.84	.71	.69-.74	.73
Block Design	.66-.90	.72	.80-.89	.86
Object Assembly	.56-.68	.64	.63-.73	.71
Mazes	.34-.84	.76	.62-.81	.72

[1]Adapted from Table 11, Mulcahy and Watters, 1982 p. 19.
[2]Adapted from Table 9, Wechsler, 1974, p 28.
Note: Medians are rounded to 2 decimal places.

The reliability coefficients for the NWT sample compare favorably with those of the U.S. standardization sample for the same age range. The median coefficient for the former group equals or exceeds that of the latter group for seven of the ten subtests. The minimum value is dramatically lower for the NWT sample on some subtests, but this usually reflects an aberrant value for a single age group. In both samples the Verbal Scale subtests tend to have higher median coefficients than the Performance Scale subtests.

Two notes of caution are in order regarding the calculation and interpretation of this coefficient. As stated earlier, it is a measure of internal consistency and does not provide a measure of stability of the tests over time. The opportunity to retest subjects was not available to the research team and a measure of stability is therefore unavailable to date.

The second caution is in regard to a feature of WISC-R administration procedures which may spuriously inflate split-half coefficients. The use of test discontinuation criteria and age-specific starting items are designed to avoid subject boredom on elementary items and loss of self-esteem due to repeated failure on advanced items. They also allow more efficient use of testing time. Measures of internal consistency will be biased by such procedures, however, as score agreement is forced on successive untested items, thereby raising the correlation between even- and odd-numbered half-tests. The split-half coefficients should therefore be considered as upper limits to the reliability of the subtests.

There are no indices of inter-tester reliability available for the WISC-R data from NWT. Scoring reliability was sought through repeated examination of the scoring by several members of the Mulcahy and Watters research team. While in the NWT, testers exchanged several scored protocols and checked for errors in scoring or summing scores. Disagreements on the appropriate score for ambiguous responses were noted on the protocol. All WISC-R protocols were examined and rescored by Mulcahy upon the return of the testers to Edmonton.

Wechsler (1974) does not provide inter-tester reliability measures for the U.S. standardization. It should be noted that blind scoring of the WISC-R protocol would not give a completely accurate indication of the score that the second tester would have obtained by testing that child himself. Without the power to make decisions regarding the use of prompts, the blind scorer is placed in an artificial situation. With this difficulty in mind, the reader may consider one study in which inter-rater agreement on ambiguous responses was examined (Sattler, Andres, Squire, Wisely, & Maloy, 1978). Sattler et al. collected 11 ambiguous responses on each of 66 items from the Similarities, Comprehension, and Vocabulary subtests, formed 11 test forms containing one response to each item, and distributed each form to ten scorers of various levels of professional experience. Tallies were made of the number of items on which all 10 raters agreed and these tallies were averaged across groups. Total agreement was reached on only 13% of the items; 80% agreement on 44% of the items; and 96% agreement on 50% of the items. Practicing psychologists and graduate students in a testing course did not differ in agreement ratios. As Sattler et al. warn, these low rates cannot be interpreted as inter-rater reliability indices for those subtests, as a tester rarely, if ever, is confronted with a protocol containing only ambiguous responses. However, these findings illustrate the subjectivity and judgement demanded of the tester. In a cross-cultural testing situation where the subject is examined in his/her second language, the possibility of poor communication and resulting unreliable scoring is enhanced. The input of NWT psychologists to the testing and scoring procedures of the norming study prepared testers for some of the responses which only made sense in the context of knowledge about lifestyles and events in those villages. The Discussion section of this report will deal with some of the responses that the author encountered while testing children in the Keewatin District which prompted reexamination of the quantitative findings.

Validity

Measures of predictive validity for the WISC-R are not available for the NWT sample at present. There is currently no standardized achievement testing practised across the territory and grading standards are not comparable across schools. The small samples collected from some schools would render within-school regression research meaningless. Data has been collected on variables such as days absent from school, which is on a common scale across schools, and the relationship of these variables to WISC-R subtest scores is planned. The essence of the predictive validity results for U.S. and southern Canadian samples was presented in Chapter II.B.

C. Data Collection Procedure

Testers

The tests were administered by nine testers (four male, five female) who were either certified psychologists or graduate students in Educational Psychology. All testers had completed at least one graduate level course in psychological testing with practicum experience. Four teams of two testers visited three settlements each, while the remaining tester administered all the tests in one of the larger settlements. Prior to the collection of data, NWT psychologists A. Langford and B. Watters provided orientation to patterns and problems of testing in the arctic.

Test Administration

The administration procedures provided in the WISC-R manual (Wechsler, 1974) were followed. Some items were replaced with items reflecting Canadian content or content which is familiar to NWT residents. These changes are listed in Mulcahy and Watters (1982). The schools provided space for the test administration. This space was usually on school property and the majority of children were tested during school hours. Although school staff cooperated in arranging the most quiet and private testing conditions available, space and resource limitations occasionally made it difficult to test without distractions. However, the subtest order within the WISC-R and the test sequencing described below appeared to be effective in gaining and maintaining rapport with most children.

It was noted in Chapter I that the Bender Motor-Visual Gestalt Test (Koppitz, 1963) and the Goodenough-Harris Draw-A-Man Test (DAM) (Harris, 1963) were included in the Mulcahy & Watters (1982) norming project. The Bender was the first test administered, as this was expected to enhance rapport between the tester and child. Indeed, most children appeared to enjoy drawing the designs required by the test. The WISC-R was then administered to the child. If the child appeared to be having a great deal of difficulty understanding the tester's conversation or instructions due to hearing difficulties or a lack of familiarity with English, the tester would note this on the test form. Many of these test profiles were eliminated from analysis, as testing these children for clinical purposes would have been inappropriate. In severe cases the tester would select a child from the pool of alternate subjects.

The DAM was administered as a group test by classroom teachers, according to standard instructions (Harris, 1963) provided them by the tester in their village. Subjects who were absent from school on the day of DAM administration were either tested by the classroom teacher upon returning to school or by the visiting tester following administration of the WISC-R.

Testers attempted to test as many of the original 50 children from each age group as possible. Tests were occasionally administered after school hours to include children who were frequently absent from school. In total, 398 WISC-R and Bender score profiles were collected for the eight age groups. Mulcahy and Watters (1982) report that 32 of these profiles were deleted when further examination of the WISC-R protocols led these authors to determine that the children in question did not have sufficient English language skills to permit valid assessment with that test. Specifically, eleven children were excluded from the 7 year age sample; nine children at 8 years; five at 9 years; two at 10 years; one at 11 years; three at twelve years; and one from the 14 year age sample. This left the total sample of 366 children which is described in Table 1.

The raw scores on each subtest were standardized for each of the eight age years and rescaled to have means of 10.0 and standard deviations of 3.0.

D. Statistical Analysis Procedure

The calculations, tests of hypotheses, and decision rules employed in producing the results reported in Chapter IV are described in the following section. The present study concerns the factor structure inherent in the scaled subtest scores and all the analysis pertains to these scores. Readers interested in the scaling procedures are referred to Mulcahy & Watters (1982). The computer programs used in the analysis are described first, followed by an explanation of statistical notation used throughout the remainder of the thesis. Testing of the assumption of normality for each of the variables and calculation of the correlation matrices for each age group will be dealt with very briefly. The tests of homogeneity of covariance and correlation matrices across age groups will be described in detail, as these procedures are more heavily debated in the literature and and because they are central to much of the factor analytic procedure and discussion to follow. The method of pooling covariance matrices will be briefly noted. The bulk of this section will describe the procedures and decision rules used in the factor analysis.

Commercial Statistical Software

A number of statistical software packages were used in the analysis reported in this document. These packages are described below. Later reference to their use will simply cite the appropriate program name to avoid repetitive interruptions in the flow of text.

The following programs are contained in the XDER library of programs maintained by the Division of Educational Research Services (DERS) at the University of Alberta.

1. DEST02: The covariance and correlation matrices for each age year were calculated from the scaled subtest scores and stored on disk by this program.

2. FACT20: The principal components and principal factor analyses were calculated with this program.
3. MULV58: A test of the equality of two or more correlation matrices is provided, with a χ^2 goodness-of-fit test derived by Jennrich (1970).

The Statistical Package for the Social Sciences (Hull & Nie, 1981; Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975) is distributed by McGraw-Hill Book Co. and is available through Computing Services at the University of Alberta. A number of nonparametric tests were performed via SPSS subprograms.

LISREL was used for all the confirmatory and exploratory maximum likelihood factor analyses in the present study. The program was written by K. Jöreskog and D. Sörbom of the University of Uppsala, Sweden and combines many of their earlier maximum likelihood programs for factor analysis, path analysis, and other statistical procedures. Version IV (1978) and Version V (1981) are available through Computing Services at the University of Alberta. Version IV was used for most of the analysis reported in this document. Version V became available as the analysis was nearing completion and was used principally to check the results of Version IV.

Pooling of the covariance matrices and the calculation of pooled correlation matrices were performed by an unpublished program written in APL by the present author. All data transformations and analyses were performed on the

University of Alberta's Amdahl 470V/8 and its recent replacement, the Amdahl 5860.

Statistical Notation

The statistical notation refers primarily to the four matrices which define the factor analytic model presented in Equation II.7.

$$\Sigma = \Lambda\Phi\Lambda' + \Psi^2$$

where Σ refers to the population covariance matrix of observed variables; Λ to the factor pattern matrix; Φ is the matrix of covariances among factors; and Ψ is the covariance matrix for the test-specific component of an observed score. Σ will be a correlation matrix in this analysis unless otherwise specified. Φ will be fixed to be a correlation matrix. Ψ is generally assumed to be diagonal, with the diagonal elements defined as the unique variance for the corresponding test. The sample covariance matrix will be defined as S; the sample correlation matrix as R. Where a subscript is used with a matrix symbol, (Λ_k), k refers to the age pool represented. For example, S_{7-8} refers to the pooled test covariance matrix for the 7 and 8 year old samples. This subscript will generally be used only when comparing parameters across groups.

Elements within the above matrices are defined as follows:

λ_{im} refers to the loading of test i on factor m, i.e., the element in row i and column m of Λ.

ϕ_{mn} refers to the correlation between factors m and n, i.e., the element in row m and column n of Φ.

ψ_{ij} refers to the covariance of residual terms for tests i and j, i.e., the element in row i and column j of Ψ.

$\psi_{2,2}$ is the unique variance for Test 2. Individual sample correlations and covariances will be represented as r_{ij} and s_{ij}, respectively.

The LISREL Model

LISREL is designed to estimate parameters within structural equations. It generally assumes a causal structure among a set of latent variables or hypothetical constructs, some of which are independent variables; others, dependent variables. These hypothetical constructs (or factors) are also the underlying cause of the observed variables (tests). Three models of the relationships among these observed and latent variables are expressed in the equations (Jöreskog & Sörbom, 1978, p. 4):

$$B\eta = \Gamma\xi + \zeta \qquad \text{(III.1)}$$

$$y = \Lambda_y\eta + \epsilon \qquad \text{(III.2)}$$

$$x = \Lambda_x\xi + \delta \qquad \text{(III.3)}$$

The first of these is the structural equation model, describing the relationships among scores on the independent (ξ) and dependent (η) variables. B (named BE in the **LISREL** command language) is the matrix of covariances among dependent latent variables and is the **LISREL** eqivalent of Φ for the dependent variables. Γ (GA) is the matrix of regression coefficients describing the regression of η on ξ. ς refers to the unexplained variance in η scores. The second and third equations are the factor models for the observed dependent and independent variables, respectively, with ϵ and δ defined as the test-specific component of these variables. The factor patterns for the dependent and independent variables are Λ_y (LY) and Λ_x (LX), respectively. Readers should note that the measurement models in the second and third equations are equivalent to the factor analytic model presented in Equation II.0, but with some changes in statistical notation.

It is at this point that **LISREL** deviates from most factor analytic literature (including Jöreskog's) in naming matrices in the factor model. While Φ (named PH) is defined as the correlations among latent dependent variables, Ψ (named PS) is defined as the covariance matrix for ς. The covariance matrices for ϵ and δ are designated $\Theta\epsilon$ (TE) and $\Theta\delta$ (TD). Clarity would be served by referring to matrices by the symbols commonly found in the literature, rather than the names assigned in the **LISREL** command language. The latter is used only to describe the steps involved in specifying a factor analytic model with **LISREL**.

The **LISREL** user may specify parameters in the matrices GA, BE, LY, LX, PH, PS, TE, and TD to be fixed to some specific value or to be free (or relaxed) to be estimated by the model. For confirmatory factor analysis, only four parameters need be specified and only three are of theoretical interest. If the number of either independent or dependent variables is unspecified in the **LISREL** commands, the program assumes that no observed or latent variables of that type exist. The factor model was defined by excluding independent variables in the above manner; specifying BE as an identity matrix; fixing the diagonal elements of PS to be 1.0; and specifying TE as a diagonal matrix with free diagonal elements. Thus, the **LISREL** matrix PS corresponds to Φ; the **LISREL** matrix LY to Λ; and the **LISREL** matrix TE to Ψ. PS was specified as diagonal for orthogonal models; symmetric for oblique models. LY elements were specified as free where loadings were hypothesized to be salient and fixed at 0 otherwise. TE was specified to be diagonal, with the diagonal elements free to be estimated by the program. In isolated instances, off-diagonal elements of TE were freed. The matrices in the factor analytic model, their corresponding **LISREL** labels, and the usual designation for these parameters is summarized in Table 3.

If a significantly large χ^2 is obtained for a given model, the researcher needs to determine which parameters are incorrectly specified. **LISREL** provides several of the diagnostic indicators of the sources of distortion in a model which were mentioned in Chapter II.E. First-order derivatives of the loss function (See Equation II.9) for the fixed and constrained parameters are output by both **LISREL IV** and **LISREL V**. Large derivatives indicate that freeing the corresponding parameter would result in some improvement to

Table 3

LISREL Matrices in the Factor Analytic Model

Matrix Name	Factor Model	LISREL Label	Specification
Factor Pattern	Λ	LY	Salient variables are free; others fixed at 0.
Correlations Among Factors	Φ	PS	Diagonal elements fixed at 1.0; Off-diagonals at 0 for orthogonal models, free for oblique models.
Unique Variance and Covariance	Ψ	TE	Diagonals free; Off-diagonals usually fixed at zero.

the model's fit to the data. **LISREL V** provides modification indices which are calculated as the ratio of the squared first-order derivative to the second-order derivative for that parameter. Jöreskog (1981) recommends the use of the modification indices rather than the first-order derivatives, as the latter are affected by the magnitude of the data and the parameter values. The present author compared modification indices output by **LISREL V** for several models to the corresponding first-order derivatives and found that decisions made on the basis of the latter index corresponded closely to those from the former index.

The matrix of residual variances and covariances is provided in the output of both **LISREL IV** and **LISREL V**. The latter program also prints a matrix of normalized residuals. Jöreskog (1981) suggests that normalized residual coefficients greater than 2.0 indicate that the corresponding sample variance-covariance coefficient is not adequately explained by the factor model under investigation.

The parameter specifications required to test the major hypotheses stated in Chapter II.F are described in the remaining sections of this chapter. The sequences of relaxing and fixing parameters for model modification are described for the total sample analyses in Chapter IV and for the age pool analyses in Appendix 1.

Deriving and Testing the Correlation Matrices

The scaled scores for each age group on each of the twelve subtests, the Verbal, Performance, and Full Scales were stored on disk in the Amdahl computer. The mean and moments about the mean were calculated for each variable by the SPSS subprogram FREQUENCIES. The assumption of normality of subtest scores was the focus of Hypothesis 1 in Chapter II.F. A normal distribution with a mean of 10.0 and standard deviation of 3.0 was specified for each subtest at each age level. The Kolmogorov-Smirnov one-sample test statistic (Massey, 1951; Smirnov, 1948) was calculated by the SPSS subprogram NPAR TESTS to examine the fit of each variable to this specification. The skewness and kurtosis figures from the FREQUENCIES results were examined for clues on the nature of any violations of the assumptions.

The XDER program DEST02 calculated 12x12 matrices of covariance and correlations among subtests for each age group. The equality of these matrices across age groups was tested in a number of ways:

1. Jennrich (1970) describes a test of the equality of two or more correlation matrices, yielding a χ^2 statistic of the goodness of fit. The degrees of freedom are calculated as $(k-1)p(p-1)/2$, where k is equal to the number of matrices compared and p is equal to the number of variables. The XDER program MULV58 was used to apply this statistical test to the set of eight correlation matrices. As detailed in Chapter IV, the hypothesis of equality of correlation matrices was not rejected for the full set of eight age groups, as specified in Hypothesis 2; or for the four pairs of age groups as specified in Hypothesis 3.
2. LISREL was used to test the equality of both the covariance and correlation matrices for the full set of eight age groups. This application is simply a special case of Jöreskog's SIFASP model, which was described in Chapter II.D. It involved specifying the factor pattern matrix as a 12x12 Identity matrix, ($\Lambda=I$); the error matrix as a zero matrix, ($\Psi=0$); and allowing the matrix of correlations among factors (Φ) to represent the sample covariance matrix. The Φ_7 matrix was free to be estimated by the minimizing function of the program, while estimates for every other Φ_k matrix were constrained to be equal to Φ_7. The sample covariance matrix for each age group is read by the program and

maximum likelihood estimates of ϕ_{ij} are calculated. Program output includes a χ^2 test statistic for the fit of the estimated covariance matrix to all eight sample covariance matrices examined simultaneously. The degrees of freedom are calculated by summing the degrees of freedom associated with the covariance matrix of each group, This is calculated as $p(p+1)/2-t$, where t is the number of free parameters in the model for that group. The hypothesis of equality of covariance matrices was not rejected with this test.

3. The above SIFASP test was conducted with the set of eight correlation matrices in place of the covariance matrices. The same decision was reached. Similarities among the χ^2 values for the above sets of tests are noted in Chapter IV and discussed in terms of the legitimacy of analyzing the correlation matrix.

The above results, for which details are provided in Chapter IV, provide justification for the pooling of covariance matrices of two or more samples. These pooled matrices were calculated as

$$S=\sum_{k=1}^{k}(N_k-1)S_k \;/\; \sum_{k=1}^{k}(N_k-1) \qquad \text{(III.4)}$$

where S_k and N_k are the covariance matrix and sample size, respectively, for each group k within the K groups pooled. Five pairs of pooled covariance and correlation matrices were calculated, according to the rationale provided in

Chapter II.E. The new matrices were the result of pooling those of ages 7 and 8; 9 and 10; 11 and 12; 13 and 14; and all eight matrices to form total sample covariance and correlation matrices. The sample sizes associated with these matrices are 71, 97, 102, 96, and 366, respectively. The factor analysis was conducted on the pooled correlation matrices. Before describing the steps in that analysis, clarification is required on two isssues.

The first issue for clarification is semantic. The term "age group(s)" has referred to the samples of children within a one-year age range, e.g. 7 years, o months to 7 years, 11 months. This designation will be continued throughout the thesis. Pooled samples of two age groups, e.g. the pooled 7 and 8 year old samples, will be referred to as "age pools". The term "total sample" will continue to refer to the 366 subjects across all eight age groups.

The second issue for clarification concerns the analysis of correlation matrices rather than covariance matrices. As stated in Chapter II.D, Jöreskog (1979a) has stated that the sample covariance matrix to be analyzed "may be taken to be a correlation matrix if the model is scale free and if the units of measurements in the variables are arbitrary or irrelevant"(p. 46). The scaled scores fit the above criteria, having been scaled to identical means and standard deviations for ease of comparison. The covariances of raw scores are not of interest in this research, since relationships among raw scores are not clinically interpreted. The correlation matrix is therefore a legitimate object of analysis and decisions reached by this analysis should be identical to those reached by analysis of the covariance matrices. In pilot experimentation with **LISREL**, this argument was empirically confirmed with a restricted three-factor solution. Analysis of the covariance matrix resulted in loadings of a larger scale but the value of χ^2, the probability of χ^2 under the null hypothesis, and the t-values for free parameters were identical for the two modes of analysis. (This result does not generalize to tests of the homogeneity of covariance matrices, in which the elements of the covariance matrix are directly fixed or freed. This qualification is discussed in Chapter IV in regard to results for the **MULV58** and **LISREL** tests described above.) The scale of loadings estimated by analysis of the correlation matrix are more easily interpreted and compared to the results of other factor analytic methods or studies. Therefore, this matrix was analyzed according to the following steps.

Factor Analysis Sequence

Fit to the Clinical Models

The following sequence of procedures was performed on the correlation matrices of each of the age pools and the total sample to test the fit of the various clinical models and to discover any required modifications to those models.

Principal Components and Factors

Principal components and principal factor analysis were conducted on each of the pooled matrices to allow comparison of results across factoring methods and provide initial estimates for the free parameters. Up to six principal components were extracted from the matrix and the number of components with eigenvalues greater than 1.0 was noted. Kaiser (1970) has suggested this value as an index of the number of common factors to interpret. This value was later compared to the number of factors suggested by maximum likelihood analysis. Principal factor analysis was conducted, with squared multiple correlations replacing the 1.0's in the diagonal of the correlation matrix. Varimax and Promax rotations were performed on up to six factors. Pattern matrices and correlations among factors are presented in Chapter IV for the total sample and in Appendix 1 for the four age pools.

Maximum Likelihood Factor Analysis

The ML analysis employed a mixture of confirmatory and exploratory techniques. An unrestricted ML solution was tested for the sufficiency of each of one, two, three, and four factors. The clinical model associated with a given number of factors was tested, even when the result for the unrestricted solution indicated that more factors were required. This is not an instance of discarding unwanted results but an accomodation to Bentler and Bonett's (1980) contention that it is not impossible for a very restricted model to achieve a good fit where a less restricted model has failed. This did not occur in any of the present analyses. Subsequent analysis to modify a poor-fitting clinical model was not pursued when the number of factors associated with that model was insufficient (except in regards to some particular hypotheses about Digit Span and Mazes, to be detailed in Chapter IV).

The intial estimates for Λ were the factor loadings from the rotated pattern matrix of the same order from the principal factor analysis for that group. The initial estimates for Φ were the correlations among M factors as reported for the appropriate principal factor analysis (as defined for Λ above). The diagonal elements of Ψ were always free and initial estimates were calculated as $1-h^2$, where h^2 refers to the communalities reported for the appropriate principal factor analysis.

The choice of parameters to free or fix involved examination of the first derivatives for fixed parameters, the confidence intervals around free parameters, the original correlation matrix, the residual correlation matrix, and theoretical hints as to possible errors in the model. This exploratory model modification was continued until the incremental fit of freeing a single parameter was nonsignificant. This criterion was preferred over the criterion of achieving a nonsignificant χ^2 for the modified model because the relatively small sample sizes of the age pools raised concerns about the the power of the χ^2 test.

The parameter specifications for Hypotheses 4 to 14 are described below.

1. The sufficiency of a single general factor was tested. This required specifying Λ as a column vector with one fixed non-zero loading and Φ as a diagonal matrix with the single diagonal element fixed to 1.0 (in effect, fixing Φ to be the scalar 1.0).
2. The sufficiency of two factors (H_0: M=2) was tested for the orthogonal case. An unrestricted model was constructed by fixing only four parameters (the square of the number of factors) in the common factor space. The two diagonal elements of Φ were fixed to 1.0 and the off-diagonal element to 0. One element of Λ was fixed to 0 ($\lambda_{1,2}$). These conditions are modeled after Jöreskog (1969, 1979b) to meet the requirements for a unique unrestricted solution.
3. The fit of the Verbal and Performance Scales as two factors was tested. An orthogonal model with simple structure was tested first. This model is diagrammed in Table 4, with 1's and 0's indicating that the parameters are fixed to the displayed value and Greek symbols indicating free parameters. As the χ^2 value indicated a significantly poor fit for all samples, all off-diagonal elements of Φ were relaxed for each sample to test the oblique model with simple structure. The modification indices almost invariably pointed to the fixed correlations among factors as the greatest source of

distortion in the orthogonal case (regardless of the number of factors extracted). If two factors were sufficient according to the fit of the unrestricted solution, parameters were freed one at a time and the incremental improvement tested by calculating the difference in χ^2 values for the successive models and comparing this value to the appropriate degrees of freedom.

4. The sufficiency of three factors was tested with the orthogonal unrestricted solution as in the two-factor model. All diagonal elements of Φ were fixed to 1.0 and the three off-diagonal elements fixed at 0. All but 3 elements of Λ were free, with $\lambda_{1,2}$, $\lambda_{1,3}$, and $\lambda_{2,3}$ fixed to 0. This procedure fixed $M^2=9$ parameters in the common factor space, as necessary for a unique solution. The choice of Λ elements to fix was somewhat arbitrary, although the selection avoided linear dependencies among the columns of Λ (Jöreskog, 1969).
5. Kaufman's three-factor model was tested for both the orthogonal and oblique cases. The oblique model is diagrammed in Table 5. Modifications were made to the model as necessary, using the rules described for the Verbal-Performance model in 3. above.
6. The sufficiency of four factors was tested for the orthogonal case. As in the other unrestricted solutions, M^2 fixed parameters were required for a unique solution, where M is the number of factors. Φ was an identity

Table 4

Clinical Models to be Tested
Wechsler Verbal and Performance Scales as Two Factors'

TEST	V	P	Ψ
Inf.	λ	0	ψ
Sim.	λ	0	ψ
Ari.	λ	0	ψ
Voc.	λ	0	ψ
Com.	λ	0	ψ
D.S.	λ	0	ψ
P.C.	0	λ	ψ
P.A.	0	λ	ψ
B.D.	0	λ	ψ
O.A.	0	λ	ψ
Cod.	0	λ	ψ
Maz.	0	λ	ψ

Intercorrelations Among Factors

	V	P
V	1	
P	0	1

'Orthogonal model
NOTE: Greek letters signify free parameters; fixed parameters are represented by their specified values.

Table 5

Clinical Models to be Tested
Kaufman Three-factor Model'

TEST	VC	PO	FD	ψ
Inf.	λ	0	0	ψ
Sim.	λ	0	0	ψ
Ari.	0	0	λ	ψ
Voc.	λ	0	0	ψ
Com.	λ	0	0	ψ
D.S.	0	0	λ	ψ
P.C.	0	λ	0	ψ
P.A.	0	λ	0	ψ
B.D.	0	λ	0	ψ
O.A.	0	λ	0	ψ
Cod.	0	0	λ	ψ
Maz.	0	λ	0	ψ

Intercorrelations Among Factors

	VC	PO	FD
VC	1		
PO	φ	1	
Seq	φ	φ	1

'Oblique model
NOTE: Greek letters signify free parameters; fixed parameters are represented by their specified value.

matrix and six elements in Λ were fixed to zero.

7. Bannatyne's four factor model was tested and modified according to the general rules outlined in 3. above. Table 6 displays the fixed and free parameters for the oblique model. The-four factor solution was only tested in the total group and the 13-14 age pool. Convergence to a minimum of F (the loss function, Equation II.9) was not achieved in either case and no interpretable solution was evident in the test or modification of Bannatyne's four-factor model. Specific findings which led to the abandonment of this model (or more precisely, the fourth factor in the model) are detailed in Chapter IV.

Cross-validation of Total Sample Results

A factor pattern derived by modifying Kaufman's three-factor model was the only acceptable and interpretable pattern obtained for the total sample from the models tested. This modified factor model was tested against the correlation matrices for the four age pools. Where the modification of various clinical models introduced an exploratory mode to the analysis, this final procedure reintroduced confirmatory analysis and allowed investigation of the generalizability of the results for the total sample. Generalization to an age pool sample would be validated by a nonsignificant χ^2 value and significant nonzero values for free parameters.

Table 6

Clinical Models to be Tested
Bannatyne Four-factor Model'

TEST	Con	Sp	Seq	AK	Ψ
Inf.	0	0	0	λ	ψ
Sim.	λ	0	0	0	ψ
Ari.	0	0	λ	λ	ψ
Voc.	λ	0	0	λ	ψ
Com.	λ	0	0	0	ψ
D.S.	0	0	λ	0	ψ
P.C.	0	λ	0	0	ψ
P.A.	0	0	λ	0	ψ
B.D.	0	λ	0	0	ψ
O.A.	0	λ	0	0	ψ
Cod.	0	0	λ	0	ψ
Maz.	0	λ	0	0	ψ

Intercorrelations Among Factors

	Con	Sp	Seq	AK
CON	1			
Sp	φ	1		
Seq	φ	φ	1	
AK	φ	φ	φ	1

'Oblique model
NOTE: Greek letters signify free parameters; fixed parameters are represented by their specified value.

IV. Results

The presentation of the results follows the order of analysis described in Chapter III.D. The findings regarding the normality distribution of subtest scores are followed in Section A by evidence of the equality of covariance and correlation matrices across age groups. The pooled correlation matrices for the total sample and the four age pools complete this section. The results of the confirmatory and exploratory factor analysis for these samples are provided in Section B, with an examination of similarities and trends among the various samples. The results of additional analyses, which were conducted on Digit Span to test hypotheses generated by the results of Section B and clinical observations of the author, are described in Section C.

A. Tests of Assumptions Regarding the Data

Normal Distribution of Scaled Scores

The Kolmogorov-Smirnov one-sample test provides an index of the fit of the scaled subtest scores to the assumption of a normal distribution with a mean of 10 and standard deviation of 3. The Kolmogorov-Smirnov Z (Smirnov, 1948) and its two-tailed probability under the null hypothesis of normality is calculated by the SPSS subprogram NPAR TESTS. Table 7 lists those subtests with distributions which differed significantly from normality at each of the

Table 7

Subtests with Distributions Deviating from N(10,3)
Kolmogorov-Smirnov One-sample Test
Age Group by Significance Level

	Level of Significance			
Age	.01	.05	.10	.25
7		Inf. Sim.	D.S. Maz.	Com.
8	Sim.	B.D.		Ari. Inf. D.S. Maz.
9		Sim.	Com. D.S. Cod.	Inf. P.C. B.D.
10		Inf. P.C. O.A.	Ari.	Sim. D.S. Com.
11	Ari.		Sim. Com. O.A. Maz.	Voc. P.A.
12	D.S.	P.A. O.A.	Inf. Sim. P.C.	Ari. Com.
13		Ari. D.S. O.A. Cod.		Inf. Com. Maz.
14	P.C.	Inf.	O.A.	Sim. Ari. D.S. Maz.

various age levels. Four possible critical levels of Type I error (.01, .05, .10, .25) are tabulated and subtests are listed under the most conservative level at which the hypothesis of normality would be rejected. An examination of histograms, kurtosis and skewness indices provided by the SPSS subprogram FREQUENCIES suggests that subtest distributions which differ from normality at the .10 level were difficult to distinguish from those of subtests which do not appear on the list. Subtest distributions which differ from normality at the .05 level were easily identified as skewed and/or peaked or flat. The following discussion is largely limited to subtests which were identified at the .05 and .01 levels of significance.

The number of subtest distributions deviating from normality at the .25 level ranged from five to eight across the age groups, while the number which are significant at .05 ranged from one to four. Certain subtests appeared more frequently than others. Similarities and Information were each listed at the .01 or .05 level for three age groups and at the .25 level at seven ages. At the other extreme, Block Design appeared only once at the .05 level, with a slightly flat distribution for the 8 year age group, and once at the .25 level. Identification of the factors underlying tests such as Similarities and Information must be viewed with extra caution considering these frequent violations of the assumption of a normal distribution. The determination of the extent to which the results of this study are affected awaits further research on the robustness of LISREL and other maximum likelihood methods to such violations.

The most striking feature of the qualitative nature of such violations was the tendency for Verbal Scale subtests to have postively-skewed distributions in contrast to the negative skewness of several of the Performance Scale subtests. Information and Similarities were particularly consistent in this regard, as were Picture Completion and Object Assembly. Arithmetic tended to be negatively skewed when nonnormal. Although this is a Verbal Scale subtest, the main analysis indicated that it loaded on both factors in the two-factor solutions for age pools 9-10 and 11-12 years and for the total sample. These trends in the shape of subtest distributions are consistent with the literature cited in Chapter II, which reports high Spatial and low Verbal scores for Inuit and Indian children. There are statistical and clinical implications for such trends. The size of correlations between pairs of Verbal and Performance subtests will be restricted by these differences. Consequently, it is difficult to determine whether the appearance of separate "Verbal" and "Spatial" factors is a result of separate underlying cognitive processes or differences in the difficulty levels of the two sets of tests. Clinically, comparison of relative strengths and weaknesses is made hazardous by the fact that a given scaled score may represent different percentiles across tests at a given age level. These concerns will be discussed in more depth in Chapter V.

Equality of Covariance and Correlation Matrices

The validity of pooling covariance matrices across age groups was tested with Jennrich's (1970) test of the equality of correlation matrices and simultaneous analysis of correlation and covariance matrices across age groups with **LISREL**. The results of these tests are in Table 8. The hypotheses of equality of covariance and correlation matrices were not rejected in any case.

Jennrich's test appears to be more powerful than simultaneous analysis of correlation matrices with **LISREL**, as the Type I error probability is much lower for the former test when all eight groups are compared. The pooling of all covariance matrices is supported by the nonsignificant χ^2 results. The use of two-year age pools for independent verification of the total sample results was supported by the absence of significant χ^2 values for matrix comparisons within age pools.

The SIFASP technique of comparing matrices, allowed by **LISREL**, provides additional support for the pooling of covariance matrices across age groups. It should be noted that the analysis of covariance matrices appears to be more rigorous than the analysis of correlation matrices. This result is in contrast to the testing of a restricted factor model for a single sample, as described in Chapter III, in which the value of χ^2 and the t-values for the various free

Table 8

Analysis of Equality of Covariance and Correlation Matrices
Summary of MULV58 and LISREL Results

Matrices Compared	χ^2	d.f.	Prob.
MULV58 Results			
Correlations: All eight age groups	472.28	462	.367
Correlations: Ages 7, 8 years	62.21	66	.609
Correlations: Ages 9, 10 years	48.68	66	.946
Correlations: Ages 11, 12 years	75.64	66	.195
Correlations: Ages 13, 14 years	65.56	66	.492
LISREL Results			
Correlations, σ^2 fixed: All groups	546.99	558	.622
Correlations, σ^2 free: All groups	546.99	546	.480
Covariances, σ^2 fixed: All groups	563.87	558	.423
Covariances, σ^2 free: All groups	561.19	546	.317

parameters were independent of the type of matrix used. This discrepancy suggests that while the analysis of correlation matrices appears to be appropriate for the testing of restricted factor models on a single sample, the simultaneous analysis of matrices via LISREL should be carried out on the covariance matrices. The main analysis for this study involved only the former case. For this reason and the rationale presented in Chapter III, correlation matrices were used for the confirmatory factor analysis of each of the age pools' data. These pooled correlation matrices were derived according to the method cited in Chapter III and are presented in Tables 9 to 13.

Table 9

Subtest Intercorrelations for Age Pool 7-8

Test	Inf.	Sim.	Ari.	Voc.	Com.	D.S.	P.C.	P.A.	B.D.	O.A.	Cod.
Sim.	.591										
Ari.	.599	.490									
Voc.	.717	.610	.624								
Com.	.528	.549	.522	.747							
D.S.	.403	.409	.520	.553	.406						
P.C.	.290	.322	.406	.317	.293	.273					
P.A.	.506	.483	.463	.534	.480	.436	.498				
B.D.	.134	.291	.215	.236	.155	.230	.543	.405			
O.A.	.166	.160	.268	.211	.144	.412	.334	.394	.445		
Cod.	.318	.125	.342	.177	.181	.169	.319	.240	.236	.204	
Maz.	.352	.400	.393	.480	.376	.248	.242	.267	.172	.312	.200

Table 10

Subtest Intercorrelations for Age Pool 9-10

Test	Inf.	Sim.	Ari.	Voc.	Com.	D.S.	P.C.	P.A.	B.D.	O.A.	Cod.
Sim.	.588										
Ari.	.523	.481									
Voc.	.724	.597	.525								
Com.	.694	.648	.534	.713							
D.S.	.323	.260	.381	.203	.146						
P.C.	.298	.433	.287	.357	.316	.128					
P.A.	.340	.385	.321	.379	.371	.364	.286				
B.D.	.364	.403	.386	.264	.261	.452	.237	.420			
O.A.	.347	.303	.271	.265	.302	.460	.276	.383	.473		
Cod.	.430	.487	.493	.412	.437	.441	.165	.358	.486	.326	
Maz.	.295	.367	.469	.327	.275	.327	.198	.272	.367	.152	.413

Table 11

Subtest Intercorrelations for Age Pool 11-12

Test	Inf.	Sim.	Ari.	Voc.	Com.	D.S.	P.C.	P.A.	B.D.	O.A.	Cod.
Sim.	.595										
Ari.	.492	.491									
Voc.	.673	.711	.509								
Com.	.532	.477	.343	.645							
D.S.	.183	.286	.469	.234	.124						
P.C.	.151	.346	.295	.213	.098	.222					
P.A.	.194	.172	.289	.194	.102	.297	.418				
B.D.	.207	.323	.355	.197	.175	.309	.321	.341			
O.A.	.211	.196	.252	.264	.119	.214	.459	.456	.493		
Cod.	.087	.123	.271	.160	.130	.369	-.017	.276	.322	.284	
Maz.	.201	.196	.205	.164	.161	.390	.161	.179	.280	.254	.336

Table 12

Subtest Intercorrelations for Age Pool 13-14

Test	Inf.	Sim.	Ari.	Voc.	Com.	D.S.	P.C.	P.A.	B.D.	O.A.	Cod.
Sim.	.481										
Ari.	.475	.262									
Voc.	.678	.687	.418								
Com.	.564	.610	.274	.696							
D.S.	.328	.179	.394	.253	.107						
P.C.	.282	.345	.152	.326	.387	.153					
P.A.	.244	.383	.187	.373	.396	.096	.355				
B.D.	.291	.320	.280	.355	.209	.285	.447	.308			
O.A.	.138	.237	.088	.095	.142	.008	.327	.161	.469		
Cod.	.239	.099	.314	.212	.132	.160	.171	.086	.272	.187	
Maz.	.125	.053	.028	.104	.055	.154	.303	.164	.331	.062	.111

Table 13

Subtest Intercorrelations for Total Sample

Test	Inf.	Sim.	Ari.	Voc.	Com.	D.S.	P.C.	P.A.	B.D.	O.A.	Cod.
Sim.	.563										
Ari.	.517	.429									
Voc.	.696	.654	.510								
Com.	.583	.573	.409	.697							
D.S.	.299	.275	.438	.292	.180						
P.C.	.250	.364	.280	.300	.267	.192					
P.A.	.307	.345	.304	.357	.327	.289	.383				
B.D.	.254	.336	.316	.263	.202	.323	.375	.364			
O.A.	.218	.229	.216	.207	.178	.260	.353	.344	.472		
Cod.	.263	.213	.352	.243	.221	.294	.146	.240	.332	.252	
Maz.	.233	.240	.259	.248	.201	.285	.222	.215	.296	.184	.268

B. Factor Analysis Results

The results for the total sample are described first, followed by those of age pools 7-8, 9-10, 11-12, and 13-14 years. A table summarizing the maximum likelihood analyses for a sample are followed by examination of the best-fitting two- and three-factor models. The Promax rotation of the principal factor solution for two and three factors is provided for each sample for comparison to its final maximum likelihood solution.

Total Sample

Principal component analysis suggested three factors, as indicated by the extraction of three components with eigenvalues greater than 1.0. The decision to interpret three factors is supported by maximum likelihood analysis, as summarized in Table 14. The one- and two-factor models were rejected as insufficient to explain the data

Table 14

Summary of Maximum Likelihood Analyses
Total Sample

Model Description	χ^2	d.f.	Prob.
1. Null Model	1508.70	66	.000
2. General Factor	315.75	54	.000
Two Factors			
3. Orthogonal unrestricted	83.83	43	.000
4. Verb., Perf. Scales, orthogonal	262.98	54	.000
5. Verb., Perf. Scales, oblique	166.15	53	.000
6. Free $\lambda_{6,2}$	135.86	52	.000
7. Free $\lambda_{3,2}$	114.02	51	.000
8. Fix $\lambda_{\text{[illegible]},1}$ *	114.98	52	.000
Three Factors			
9. Orthogonal unrestricted	29.35	33	.650
10. Kaufman, orthogonal	353.56	54	.000
11. Kaufman, oblique	104.14	51	.000
12. Free $\lambda_{3,1}$	88.39	50	.001
13. Free $\psi_{3,1}$	82.87	49	.002
14. Free $\lambda_{8,1}$; Fix $\psi_{3,1}$	81.18	49	.003
15. Free $\lambda_{12,3}$	69.54	48	.023
16. Free $\lambda_{7,1}$	62.53	47	.064
17. Free $\lambda_{2,2}$	51.18	46	.278
18. Free $\lambda_{1,3}$	43.74	45	.525
19. Fix $\lambda_{12,2}$ *	45.62	46	.488
20. Fix $\lambda_{7,1}$, $\lambda_{2,2}$, $\lambda_{1,3}$; Free $\psi_{7,2}$, $\psi_{3,1}$	58.96	47	.113
21. Free $\lambda_{7,1}$, $\lambda_{2,2}$, $\lambda_{1,3}$	39.48	44	.666
Four Factors			
22. Orthogonal unrestricted	16.40	24	.873
23. Bannatyne, orthogonal	463.75	52	.000
24. Bannatyne oblique	75.00	46	.004
25. Free $\lambda_{8,2}$	74.53	45	.004
26. Free $\lambda_{8,1}$, $\lambda_{8,4}$	67.25	43	.011
27. Free $\lambda_{12,1}$, $\lambda_{12,4}$	56.09	41	.058
28. Fix $\lambda_{12,1}$, $\lambda_{8,4}$	62.33	43	.028
29. Free $\lambda_{2,2}$; Fix $\lambda_{12,4}$ *	55.51	43	.096

* These factor patterns are presented in subsequent tables.

(χ^2_{44}=315.75, p<.001 for the former; χ^2_{43}=83.83, p<.001 for the latter). The unrestricted orthogonal solution for three factors provides a satisfactory solution (χ^2_{33}=29.35, p=.650) and a significantly better fit to the data than does the two-factor orthogonal solution (χ^2_{10}=54.18, p<.01). The four-factor unrestricted model does not improve upon the three factor solution (χ^2_{8}=12.95, .10<p<.20). Therefore, the data matrix should be well described by a three-factor model. Before describing the results for the tests on Kaufman's model, some particulars of the two-factor model are noted.

Two Factors

The rejection of Wechsler's Verbal and Performance Scales as a two-factor model is not surprising, given the insufficiency of two factors in general. This insufficiency would normally lead a researcher to abandon two-factor models and concentrate on identifying the best three-factor solution. The results of the principal factor analysis, which appear in Table 15, and the relative size of partial derivatives for fixed parameters in the oblique Verbal-Performance model suggested that Digit Span loaded on the Performance factor rather than the Verbal factor and that its Performance loading should be freed. This modification to the oblique Wechsler model significantly improved the fit (χ^2_{1}=30.29, p<.001), as did the subsequent freeing of Arithmetic on the Verbal factor. (χ^2_{1}=21.84, P<.01). These procedures resulted in a nonsignificant Verbal

Table 15

Principal Factor Analysis of WISC-R Subtests
Total Sample
Promax Solution for Two Factors

	I	II	Ψ
Inf.	.776	.012	.388
Sim.	.672	.114	.449
Ari.	.451	.292	.549
Voc.	.919	-.060	.213
Com.	.822	-.092	.400
D.S.	.091	.481	.711
P.C.	.090	.489	.703
P.A.	.155	.468	.675
B.D.	-.101	.759	.500
O.A.	-.134	.681	.621
Cod.	.066	.443	.766
Maz.	.084	.384	.809
%Common Var.	56.49	43.42	
%Total Var.	50.17	38.65	

Intercorrelations Among Factors

	I
Factor II	.559

loading for Digit Span, as measured by its confidence interval. Arithmetic loaded on both factors. Fixing Digit Span's Verbal loading to 0 did not result in significant incremental distortion ($\chi^2_1=.96$, $.30<p<.50$). Table 16 defines the Verbal-Performance model at this point in the analysis.

These findings indicate that although two factors were insufficient to explain the data, relative improvement on the Verbal-Performance model was gained by considering Digit Span as a Performance test. Researchers or clinicians who continue to interpret the Verbal and Performance Scales as

Table 16

Maximum Likelihood Analysis of WISC-R Subtests
Total Sample
Final Model for Two Factors: Estimates and Standard Errors[1]

	I	II	Ψ
Inf.	.781(.046)	.0	.390(.036)
Sim.	.740(.047)	.0	.453(.039)
Ari.	.396(.061)	.323(.065)	.590(.048)
Voc.	.891(.043)	.0	.206(.028)
Com.	.765(.046)	.0	.414(.037)
D.S.	.0	.529(.055)	.720(.060)
P.C.	.0	.536(.055)	.713(.059)
P.A.	.0	.585(.054)	.657(.057)
B.D.	.0	.663(.052)	.561(.053)
O.A.	.0	.562(.054)	.684(.058)
Cod.	.0	.480(.055)	.770(.062)
Maz.	.0	.441(.056)	.805(.064)

Intercorrelations Among Factors

	I
Factor II	.582(.047)

[1] Model 8. from Table 14; (χ^2_{52}=114.98, p=.000).

two factors must avoid interpreting Digit Span as a Verbal Subtest or including it in a composite Verbal Score. Arithmetic must also be interpreted with special caution, since it loaded on both factors. These results, combined with the unrestricted analysis above, suggest that a two-factor interpretation of the WISC-R is not justified for Arctic children. The structure of the best three-factor model is described below.

Three Factors

The Kaufman three-factor model was rejected for both the orthogonal (χ^2_{34}=353.56, p<.001) and oblique (χ^2_{31}=104.14, p<.001) cases, although the latter was a significant improvement upon the former (χ^2_3=249.42, p<.001). The Promax solution for three factors, presented in Table 17, contained a Factor I loading of .317 for Arithmetic. This loading was associated with the second-largest partial derivative of all fixed parameters in Λ. The parameter was freed, resulting in significant improvement (χ^2_1=15.75, p=.001) but not in a satisfactory model. The highest partial derivative at this point was associated with the covariance of the uniqueness component of Arithmetic and Information scores, i.e. $\psi_{3,1}$. This parameter was freed and the improvement was significant (χ^2_1=5.52, p<.05). However, the 95% confidence interval for $\psi_{3,1}$ was only .067±.064 and the parameter was again fixed to 0.

Picture Arrangement was allowed to load on the Verbal factor due to the large partial derivative associated with this loading in previous model tests. The improvement was significant (χ^2_1=7.21, p<.01). (Note that this model, numbered 14 in Table 14, is compared to Model 12 rather than Model 13, as the resetting of $\psi_{3,1}$ to 0 means that Model 14 is not a subset of Model 13). Mazes were then allowed to load on the third factor. This modification was suggested by the relatively large partial derivative in previous tests and the fact that the highest loading for Mazes in the

Table 17

Principal Factor Analysis of WISC-R Subtests
Total Sample
Promax Solution for Three Factors

	I	II	III	Ψ
Inf.	.716	-.087	.174	.382
Sim.	.686	.184	-.054	.430
Ari.	.317	-.095	.552	.476
Voc.	.897	-.032	.009	.211
Com.	.834	.025	-.117	.385
D.S.	-.043	.027	.630	.612
P.C.	.168	.612	-.141	.611
P.A.	.177	.448	.053	.654
B.D.	-.105	.590	.248	.494
O.A.	-.097	.636	.078	.587
Cod.	-.046	.057	.537	.701
Maz.	.016	.130	.359	.791
%Common Var.	48.57	26.93	24.50	
%Total Var.	46.98	26.05	23.70	

Intercorrelations Among Factors

	I	II
Factor II	.460	
Factor III	.567	.586

Promax solution was on this factor. The resulting improvement was significant (χ^2_1=11.64, p<.001), although the overall model was still unsatisfactory.

Similarities was then freed on Factor I as a result of large partial derivatives on earlier tests. The improvement in the model was significant (χ^2_1=[illegible].01, p<.01) and the overall model represented a satisfactory fit to the data matrix, as indicated by the nonsignificant χ^2 for the model (χ^2_{47}=62.53, p=.064). However, further modifications were tested to detect further necessary corrections to the

Kaufman model.

The next modification was to free the loading of Similarities on Factor II ($\chi^2_1=11.35$, $p<.001$). This loading was associated with the largest partial derivative in Λ on earlier tests, but remained fixed in favor of modifications with more theoretical support and greater agreement with the three-factor Promax solution. Like $\lambda_{7,1}$, it appears to be traceable to a correlation of .364 between Similarities and Picture Completion (see Table 13). This was Picture Completion's largest correlation with any subtest, although Similarities had higher correlations with Information, Arithmetic, Vocabulary, and Comprehension. This correlation was also reflected in large partial derivatives for $\psi_{7,2}$, the covariance of the uniqueness components of these subtests. The confidence intervals for all these parameters approached but excluded 0 when free, as tested in Models 20 and 21 in Table 14. It seems that Picture Completion and Similarities shared some variance that was not accounted for by the correlations between their respective factors. The possible psychological meaning of this relationship is discussed in Chapter V.

The loading of Mazes on Factor II became nonsignificant when that test was allowed to also load on Factor III. The former parameter was fixed without significant incremental distortion in the model's fit to the data ($\chi^2_1=1.88$, $p>.05$). It should be noted that the Factor II loading was originally significant, although the partial derivative for this parameter was not large when the loading was fixed to 0. An examination of the factor structure matrix, as opposed to the factor pattern matrix represented by Λ, indicates that Mazes has a correlation of .348 with Factor II in the Promax solution and .316 with Factor II as defined by the maximum likelihood solution in Model 19, Table 14. Mazes has correlations with Factor III of .444 and .470, respectively. Mazes is better defined by the third factor than the second, although it shares some variance with the latter.

The factor solution represented by the current modifications to Kaufman's model was accepted as the best fit to the data (χ^2_{46}=45.62, p=.488). The LISREL estimates for the model are provided in Table 18. Subsequent modifications were applied in an attempt to retain simple structure in the factor pattern matrix by accounting for the relationships between Similarities and Picture Completion and between Arithmetic and Information in Ψ rather than Λ. According to the Bentler and Bonett (1980) guidelines, Model 20 cannot be tested against previous models since it is not a subset of any previous model. However, an increase in χ^2 of 13.34 with a net increase of 1 d.f. suggests that these modifications are not justified. Model 21 may be tested against Model 19, since the former is a less-restricted subset of the latter. The freeing of $\psi_{7,2}$ and $\psi_{3,1}$ did significantly improve the overall fit of the model (χ^2_2=6.14, p<.05) but the confidence intervals for both parameters included 0 (.057±.034 and .053±.030,

Table 18

Maximum Likelihood Analysis of WISC-R Subtests
Total Sample
Final Model for Three Factors:Estimates and Standard Errors[1]

	I	II	III	Ψ
Inf.	.691(.055)	.0	.150(.056)	.396(.035)
Sim.	.656(.050)	.187(.050)	.0	.444(.038)
Ari.	.316(.062)	.0	.483(.067)	.512(.048)
Voc.	.905(.043)	.0	.0	.181(.030)
Com.	.771(.046)	.0	.0	.406(.037)
D.S.	.0	.0	.620(.057)	.615(.061)
P.C.	.159(.057)	.480(.062)	.0	.689(.058)
P.A.	.225(.056)	.453(.061)	.0	.669(.056)
B.D.	.0	.741(.055)	.0	.451(.058)
O.A.	.0	.628(.055)	.0	.606(.058)
Cod.	.0	.0	.535(.058)	.713(.063)
Maz.	.0	.0	.470(.059)	.779(.065)

Intercorrelations Among Factors

	I	II
Factor II	.367(.064)	
Factor III	.505(.063)	.673(.057)

[1] Model 19. from Table 14; (χ^2_{46}=45.62, p=.488).

respectively). Therefore, these modifications were not considered essential to the model.

Examination of Table 18 indicates a number of alterations to the factor model presented by Kaufman (1975). Some of these alterations, such as Arithmetic's strong loading on the first factor, were consistent with the results of factor analysis with the U.S. standardization sample. As noted in Chapter II.B, Kaufman (1979a) suggested that third-factor subtests such as Arithmetic should not be interpreted as such unless their scores are significantly different from the Verbal and Performance Scale scores of the child tested. The loading of Mazes on the third factor is a clear departure from the Kaufman model. Its possible psychological meaning is discussed in Chapter V. Picture Arrangement's loading on Factor I is consistent with Kaufman's (1975) results, if not the associated model. The other departures from Kaufman's model involve small but significant loadings which appear to be related to relationships between Similarities and Picture Completion and between Information and Arithmetic. These loadings ($\lambda_{7,1}$, $\lambda_{2,2}$, and $\lambda_{1,3}$ were neither the largest loadings for the respective subtests nor the largest loadings on the respective factors. They do challenge the legitimacy of describing these factors as Verbal Comprehension, Perceptual Organization, and Sequencing and of deriving and comparing factor scores based on the Kaufman model.

Four Factors

The fourth factor in the Promax solution displayed in Table 19 does not resemble Bannatyne's (1974) Acquired Knowledge factor. Only Coding had a loading above .30 and the factor accounts for only 2.82% of the common variance. The first three factors resembled Kaufman's three-factor model closely, with the exception of Mazes' loading on Factor III.

When Bannatyne's model was tested by maximum likelihood methods the fit to the data was poor for both the orthogonal (χ^2_{52}=463.75, $p<0.001$) and oblique (χ^2_{48}=74.53, $p=.004$) cases.

Table 19

Principal Factor Analysis of WISC-R Subtests
Total Sample
Promax Solution for Four Factors

	I	II	III	IV	Ψ
Inf.	.730	-.087	.172	.071	.381
Sim.	.637	.180	-.026	-.102	.467
Ari.	.222	-.087	.625	.016	.459
Voc.	.901	-.036	.009	.005	.210
Com.	.892	.020	-.155	.051	.373
D.S.	-.137	.039	.702	.035	.593
P.C.	.044	.609	-.077	-.257	.578
P.A.	.130	.450	.076	-.077	.650
B.D.	-.042	.603	.203	.141	.484
O.A.	-.022	.646	.021	.109	.574
Cod.	.078	.074	.469	.339	.664
Maz.	.021	.140	.361	.103	.791
%Common Var.	46.41	26.68	24.09	2.82	
%Total Var.	46.09	26.49	23.91	2.80	

Intercorrelations Among Factors

	I	II	III
Factor II	.456		
Factor III	.649	.577	
Factor IV	-.366	-.043	.232

The largest partial derivatives in the former model were associated with the off-diagonal elements of Φ and the oblique model was a significant improvement (χ^2_6=388.75, p<.001). The first and fourth factors had a correlation above .90, however, and cannot reasonably be defined as two separate constructs. Other evidence for this conclusion is as follows: Verbal Scale subtest loadings which were fixed at 0 for either of Factors I or IV had high partial derivatives (with the exception of Digit Span loadings);

partial derivatives for Performance Scale subtests tended to be large or small for both factors; allowing Picture Arrangement and Mazes to load on both factors resulted in a solution in which these tests had strong negative loadings on either of Factors I or IV. Fixing Mazes' loading on both of factors I and IV, fixing Picture Arrangement's loading on Factor IV, and allowing Similarities to load on Factor II resulted in a solution which had an acceptable level of Type I error (χ^2_{43}=55.51, p=.096) but little interpretive clarity. This solution is presented in Table 20. The reader should note the nonsignificant Factor IV loading for Vocabulary and the extremely large, but nonetheless nonsignificant, correlation between the first and fourth factor. Results such as these led to the conclusion that Bannatyne's four-factor model for interpretation of the WISC-R does not have empirical factor analytic support. Consequently, this model was not examined in the individual age pools.

Summary of Total Sample Results

Two factors were not sufficient to reproduce or explain the sample correlation matrix. Therefore the interpretation of the Verbal and Performance Scales as a two-factor model is not supported. Specific departures from the clinical model include Digit Span's loading on the Performance Scale factor, rather than with the Verbal subtests. The Kaufman three-factor model required modification before it could be accepted as an interpretive model for this sample. The clearest departure was the loading of Mazes on the third

Table 20

Maximum Likelihood Analysis of WISC-R Subtests
Total Sample
Final Model for Four Factors: Estimates and Standard Errors'

	I	II	III	IV	Ψ
Inf.	.0	.0	.0	.806(.052)	.350(.054)
Sim.	.644(.053)	.200(.055)	.0	.0	.430(.039)
Ari.	.0	.0	.432(.076)	.368(.071)	.481(.048)
Voc.	.586(.165)	.0	.0	.324(.163)	.201(.030)
Com.	.787(.048)	.0	.0	.0	.380(.041)
D.S.	.0	.0	.627(.060)	.0	.607(.065)
P.C.	.0	.569(.055)	.0	.0	.676(.059)
P.A.	.171(.061)	.475(.104)	.011(.109)	.0	.662(.057)
B.D.	.0	.709(.053)	.0	.0	.497(.054)
O.A.	.0	.598(.055)	.0	.0	.643(.057)
Cod.	.0	.0	.525(.059)	.0	.725(.064)
Maz.	.0	.421(.057)	.0	.0	.823(.065)

Intercorrelations Among Factors

	I	II	III
Factor II	.450(.064)		
Factor III	.433(.080)	.722(062)	
Factor IV	.920(.474)	.474(066)	.623(079)

Model 28. from Table 14; (χ^2_{43}=55.51, p=.096).

factor, rather than the second, or Perceptual Organization, factor. A large correlation between Similarities and Picture Completion seemed to be the source of small but significant loadings of Picture Completion on Factor I and Similarities on Factor II. Picture Arrangement's loading on Factor I is not included in the clinical model but has been noted in other samples in literature pertaining to the WISC-R. Information's loading on the third factor seems to be attributable to a large correlation between Arithmetic and Information. Given the source and size of this loading, it should not exert a strong influence on the interpretation of this factor. Like the aberrant loadings for Similarities and Picture Completion mentioned above, this loading and the high intercorrelations among factors lead the present author to caution against the comparison of factor scores based on a strict interpretation of Kaufman's model.

Bannatyne's four-factor model received no empirical support. There appears to be nothing to gain but confusion by attempting to define a separate Acquired Knowledge factor within the Verbal Scale. Bannatyne's three original factors of Conceptualization, Spatial, and Sequential were contained within Factors I, II, and III, respectively, of the three-factor model accepted in this analysis.

Age Pool 7-8 Years

Principal components analysis extracted two components with eigenvalues greater than 1.0, suggesting the extraction and rotation of two common factors. Maximum likelihood analysis supported this conclusion, as demonstrated in the summary of ML analysis for this age pool presented in Table 21. Although the General Factor model was clearly unsatisfactory (χ^2_{54}=96.67, $p<.001$), the unrestricted orthogonal solution for two factors provided an acceptable fit to the data (χ^2_{43}=46.91, $p=.315$). The three-factor unrestricted orthogonal solution was not a significant improvement upon two factors (χ^2_{10}=15.85, $p>.05$). It does not necessarily follow that the two-factor clinical model will

Table 21

Summary of Maximum Likelihood Analysis
Age Pool 7-8 Years

Model Description	χ^2	d.f.	Prob.
1. Null Model	1102.90	66	.000
2. General Factor	96.67	54	.000
Two Factors			
3. Orthogonal unrestricted	46.91	43	.315
4. Verb., Perf. Scales, orthogonal	98.28	54	.000
5. Verb., Perf. Scales, oblique	73.34	53	.034
6. Free $\lambda_{12,1}$	65.60	52	.097
7. Free $\lambda_{8,1}$	55.26	51	.317
8. Free $\lambda_{6,2}$	53.27	50	.350
9. Fix $\lambda_{6,2}$, $\lambda_{12,2}$ *	55.76	52	.335
10. Free $\psi_{10,6}$	48.33	51	.580
Three Factors			
11. Orthogonal unrestricted	31.06	33	.564
12. Kaufman, orthogonal	142.74	54	.000
13. Kaufman, oblique	68.95	51	.048
14. Free $\lambda_{12,1}$	60.73	50	.142
15. Free $\lambda_{8,1}$ *	50.03	49	.432
16. Free $\lambda_{11,2}$	47.47	48	.494
17. Free $\lambda_{6,1}$	47.16	47	.466
18. Fix $\lambda_{6,1}$; Free $\lambda_{3,1}$	47.44	47	.455
19. Fix $\lambda_{12,1}$, $\lambda_{3,3}$, $\lambda_{11,3}$; Free $\lambda_{10,3}$	47.61	49	.592

* These factor patterns are presented in subsequent tables.

fit the data as well as Kaufman's model. The examination of both clinical models is described below.

Two Factors

The interpretation of the Verbal and Performance Scales as a two-factor model was unsatisfactory for both the orthogonal and oblique cases, although the latter was a significant improvement upon the former ($\chi^2_1=24.94$, $p<.001$).

The Promax solution for two factors, presented in Table 22, contained high loadings on the Verbal factor for Picture Arrangement and Mazes. Both of these fixed parameters had large partial derivatives in the previous test. Freeing the Factor I loading of Mazes made a significant improvement on the model (χ^2_1=7.74, p<.01) and allowed the model to have an acceptable fit to the data (χ^2_{52}=65.60, p=.097). Freeing Picture Arrangement's loading on the first factor was also a significant improvement upon the model's fit (χ^2_1=10.34, p<.01) and this modification was retained. Digit Span was allowed to load on the Performance factor to examine the generality of the total sample result to this age pool, but this modification was inappropriate and was not retained. The Factor II loading for Mazes was nonsignificant and therefore fixed to 0. The resulting increase in χ^2 was not significant (χ^2_1=0.50, p>.05). The factor pattern for this modified solution is presented in Table 23.

The covariance between the uniqueness components of Digit Span and Object Assembly had a large partial derivative in models tested to this point. An examination of Table 9 revealed an intercorrelation of .412 between the subtests' scaled scores. Freeing $\psi_{10,6}$ resulted in a significant reduction in χ^2 of 7.44 (d.f.=1, p<.01) but the resulting χ^2 was lower than the degrees of freedom for the model. Consequently, the model displayed in Table 23 was accepted as the best two-factor representation of the data.

Table 22

Principal Factor Analysis of WISC-R Subtests
Age Pool 7-8 Years
Promax Solution for Two Factors

	I	II	Ψ
Inf.	.854	-.101	.348
Sim.	.705	.026	.484
Ari.	.674	.130	.440
Voc.	.957	-.101	.171
Com.	.815	-.105	.412
D.S.	.488	.222	.604
P.C.	.087	.663	.494
P.A.	.426	.405	.480
B.D.	-.126	.785	.468
O.A.	-.060	.694	.556
Cod.	.132	.325	.834
Maz.	.457	.116	.724
%Common Var.	66.83	33.18	
%Total Var.	55.25	27.43	

Intercorrelation Among Factors

	I
Factor II	.505

Three Factors

The Promax solution for three factors, presented in Table 24, was actually suggestive of two factors. The third factor reflects the correlation between Digit Span and Object Assembly which was discussed in regards to the two-factor solution. Picture Arrangement and Mazes load on the first factor, as they did when only two factors were rotated.

The Kaufman model was rejected for the orthogonal (χ^2_{54}=142.74, p<.001) and oblique (χ^2_{51}=68.95, p=.048) cases,

Table 23

Maximum Likelihood Analysis of WISC-R Subtests
Age Pool 7-8 Years
Final Model for Two Factors: Estimates and Standard Errors[1]

	I	II	Ψ
Inf.	.773(.103)	.0	.402(.079)
Sim.	.701(.107)	.0	.508(.094)
Ari.	.715(.107)	.0	.488(.091)
Voc.	.913(.094)	.0	.167(.051)
Com.	.772(.103)	.0	.405(.079)
D.S.	.596(.112)	.0	.645(.115)
P.C.	.0	.751(.119)	.436(.119)
P.A.	.406(.117)	.444(.125)	.464(.093)
B.D.	.0	.702(.120)	.507(.121)
O.A.	.0	.545(.125)	.703(.135)
Cod.	.0	.388(.130)	.850(.151)
Maz.	.508(.116)	.0	.742(.129)

Intercorrelations Among Factors

	I
Factor II	.483(.119)

[1] Model 9. from Table 21; (χ^2_{52}=55.76, p=.335).

although the latter was only marginally significant. Allowing Mazes to load on the first factor was a significant improvement (χ^2_1=8.22, p<.01) and resulted in an acceptable fit for the overall model. Allowing Picture Arrangement to also load on Factor I significantly improved the fit of the model (χ^2_1=10.70, p<.01) and this modification was retained. The factor pattern for this modified solution is presented in Table 25. Further modifications included allowing Coding to load on Factor II; Digit Span and Arithmetic on Factor I. As is evident from Table 21, these modifications did not improve the fit of the model. Also, the value of χ^2 was

Table 24

Principal Factor Analysis of WISC-R Subtests
Age Pool 7-8 Years
Promax Solution for Three Factors

	I	II	III	Ψ
Inf.	.848	.115	-.214	.312
Sim.	.690	.092	-.034	.481
Ari.	.654	.130	.052	.439
Voc.	.934	-.122	.070	.165
Com.	.801	-.041	-.039	.411
D.S.	.443	-.176	.517	.477
P.C.	.072	.727	-.002	.421
P.A.	.400	.337	.146	.478
B.D.	-.152	.641	.242	.459
O.A.	-.112	.137	.720	.425
Cod.	.129	.453	-.110	.780
Maz.	.433	-.022	.200	.710
%Common Var.	59.78	23.28	16.94	
%Total Var.	53.16	20.70	15.06	

Intercorrelations Among Factors

	I	II
Factor II	.454	
Factor III	.435	.549

approximately equal to the model's degrees of freedom for these tests and therefore the value of further modifications is dubious. However, these tests produced some results which cast doubt on the validity of the third factor displayed in Table 25. When Digit Span was allowed to load on Factor I, its loading on Factor III was nonsignificant. Allowing Coding to load on the second factor had the same effect on its third factor loading. Allowing Arithmetic to load on the first factor resulted in nonsignificant loadings for this subtest on both factors, although its third factor loading

Table 25

Maximum Likelihood Analysis of WISC-R Subtests
Age Pool 7-8 Years
Final Model for Three Factors:Estimates and Standard Errors[1]

	I	II	III	Ψ
Inf.	.776(.103)	.0	.0	.397(.078)
Sim.	.694(.108)	.0	.0	.519(.096)
Ari.	.0	.0	.801(.112)	.359(.104)
Voc.	.923(.094)	.0	.0	.147(.052)
Com.	.779(.103)	.0	.0	.393(.078)
D.S.	.0	.0	.658(.115)	.567(.113)
P.C.	.0	.740(.120)	.0	.452(.121)
P.A.	.402(.114)	.463(.122)	.0	.465(.094)
B.D.	.0	.697(.121)	.0	.514(.122)
O.A.	.0	.576(.124)	.0	.668(.132)
Cod.	.0	.0	.358(.126)	.872(.151)
Maz.	.453(.131)	.117(.138)	.0	.736(.128)

Intercorrelations Among Factors

	I	II
Factor II	.427(.127)	
Factor III	.853(.072)	.604(.124)

[1] Model 15. from Table 21; (χ^2_{48}=50.03, p=.432).

was large (95% confidence interval = .706±.804). Table 25 reveals that the third factor had correlations of .853 and .604 with Factors I and II, respectively. It is apparent that the third factor is not stable and should not be interpreted or used to generate factor scores in clinical practice.

Summary for Age Pool 7-8 Years

The two-factor model presented in Table 23 is the best representation of the data of the models examined. The Verbal Scale is contained within a single factor but that factor also has strong loadings for Picture Arrangement and Mazes. This raises questions about the interpretation of the Performance Scale as a nonverbal factor, which will be discussed in Chapter V.

Although a three factor model resembling Kaufman's was found to have a low χ^2 value, the third factor was not stable and added no interpretive power.

Age Pool 9-10 Years

Principal component analysis indicated that three factors were required to explain the common variance among the subtests. Maximum likelihood analysis, however, indicated that two factors were sufficient to reproduce the correlation matrix. Table 26 presents the summary of the ML analyses for this age pool. The unrestricted orthogonal solution for two factors is sufficient to describe the data (χ^2_{43}=42.58, p=.489) and the three-factor orthogonal unrestricted solution does not significantly improve upon two factors (χ^2_{10}=16.25, p>.05). The general factor model was unsatisfactory. The testing of the two- and three-factor clinical models is described below.

Two Factors

The two-factor clinical model was rejected for both the orthogonal (χ^2_{54}=141.14, p<.001) and oblique solutions (χ^2_{53}=95.06, p<.001), although the latter was a significant improvement (χ^2_1=46.07, p<.001). The Promax solution presented in Table 27 suggests the clinical model with a

Table 26

Summary of Maximum Likelihood Analysis
Age Pool 9-10 Years

Model Description	χ^2	d.f.	Prob.
1. Null Model	509.98	66	.000
2. General Factor	116.36	54	.000
Two Factors			
3. Orthogonal unrestricted	42.58	43	.489
4. Verb., Perf. Scales, orthogonal	141.14	54	.000
5. Verb., Perf. Scales, oblique	95.06	53	.000
6. Free $\lambda_{6,2}$	68.01	52	.067
7. Free $\lambda_{3,2}$	59.05	51	.205
8. Free $\lambda_{7,1}$	53.95	50	.326
9. Free $\psi_{6,1}$	48.83	49	.480
10. Fix $\lambda_{6,1}$	56.62	50	.242
11. Free $\lambda_{6,1}$; Fix $\lambda_{7,2}$	49.45	50	.496
12. Fix $\psi_{6,1}$ *	54.35	51	.348
Three Factors			
13. Orthogonal unrestricted	26.33	33	.788
14. Kaufman, orthogonal	172.98	54	.000
15. Kaufman, oblique	71.32	51	.032
16. Free $\lambda_{3,1}$	64.10	50	.087
17. Free $\lambda_{2,2}$	58.27	49	.171
18. Free $\lambda_{7,1}$	53.97	48	.257
19. Fix $\lambda_{7,2}$	55.53	49	.242
20. Free $\lambda_{12,3}$	50.70	48	.368
21. Fix $\lambda_{12,2}$ *	51.63	49	.372

* These factor patterns are presented in subsequent tables.

switch. Digit Span loaded on the Performance factor, while Picture Completion loaded on the Verbal factor. Arithmetic had moderate loadings on both factors. This pattern was also reflected in the partial derivatives of the fixed parameters in the maximum likelihood analysis. Consequently, the first modification was the freeing of Digit Span's loading on Factor II, the Performance factor. The improvement was

Table 27

Principal Factor Analysis of WISC-R Subtests
Age Pool 9-10 Years
Promax Solution for Two Factors

	I	II	Ψ
Inf.	.779	.052	.344
Sim.	.695	.126	.398
Ari.	.478	.302	.513
Voc.	.906	-.108	.281
Com.	.938	-.144	.257
D.S.	-.175	.805	.485
P.C.	.386	.097	.798
P.A.	.201	.434	.670
B.D.	-.019	.733	.479
O.A.	.002	.620	.614
Cod.	.241	.525	.520
Maz.	.190	.400	.716
%Common Var.	58.33	41.67	
%Total Var.	50.76	36.26	

	Intercorrelation Among Factors I
Factor II	.579

significant ($\chi^2_1=27.05$, $p<.001$) and the overall model attained an acceptable fit to the data ($\chi^2_{52}=68.01$, $p=.067$). The other modifications suggested above were tested to determine their importance for the model.

Freeing Arithmetic's loading on the Performance factor resulted in a significant improvement in the model's fit to the data ($\chi^2_1=8.96$, $p<.01$), as did allowing Picture Completion to load on Factor I ($\chi^2_1=5.10$, $p<.05$). The covariance between the uniqueness components of Information and Digit Span was freed, as this parameter had been associated with large partial derivatives in previous tests. The reduction in χ^2 was significant ($\chi^2_1=5.12$, $p<.05$) but the resulting covariance estimate was small and only marginally significant (95% confidence interval = .119±.108) The resulting χ^2 for the overall model was smaller than the degrees of freedom. This parameter was later reset to 0.

Digit Span's loading on the first factor was fixed at 0 to test the second factor's sufficiency to explain the subtest's common variance, but the resulting distortion was significant ($\chi^2_1=7.79$, $p<.01$) and the parameter was freed. The reader should note that this loading is negative, although examination of Table 13 reveals that this subtest did not have negative correlations with the other subtests loading on this factor.

Picture Completion's loading on the Performance factor was fixed to 0 without significant decrement in the model's fit ($\chi^2_1=0.62$, $p>.05$). Note that this model (Model 11 in Table 21) is compared to Model 9 since the simultaneous fixing of $\lambda_{8,1}$ means that Model 11 is not a subset of Model 10.

The covariance between the uniqueness components of Information and Arithmetic was reset to 0 and the resulting model is presented in Table 28. The most important departure from the clinical two-factor model is the fact that Picture Completion loads only on the Verbal factor. This subtest is presumed to measure spatial skills by all three clinical models relevant to this study. Digit Span is an optional

Table 28

Maximum Likelihood Analysis of WISC-R Subtests
Age Pool 9-10 Years

Final Model for Two Factors: Estimates and Standard Errors[1]

	I	II	Ψ
Inf.	.822(.086)	.0	.324(.060)
Sim.	.753(.090)	.0	.433(.073)
Ari.	.354(.129)	.400(.132)	.518(.082)
Voc.	.840(.085)	.0	.294(.058)
Com.	.844(.085)	.0	.288(.057)
D.S.	-.369(.165)	.929(.173)	.478(.110)
P.C.	.426(.102)	.0	.819(.121)
P.A.	.0	.574(.100)	.670(.106)
B.D.	.0	.668(.097)	.554(.093)
O.A.	.0	.565(.101)	.680(.107)
Cod.	.0	.699(.096)	.511(.089)
Maz.	.0	.530(.102)	.719(.111)

Intercorrelations Among Factors

	I
Factor II	.696(.077)

[1] Model 12. from Table 26; (χ^2_{51}=54.35, p=.348).

test but its loading on the Performance factor and negative loading on the Verbal factor invalidate the calculation of Verbal factor or IQ scores which include Digit Span.

Three Factors

The Kaufman model was rejected for both the orthogonal and oblique cases, although the latter improved upon the fit of the former (χ^2_3=101.66, p<.001). The Promax solution in Table 29 suggests several required modifications to the clinical model tested. Picture Completion loads on Factor I;

Table 29

Principal Factor Analysis of WISC-R Subtests
Age Pool 9-10 Years
Promax Solution for Three Factors

	I	II	III	Ψ
Inf.	.735	.071	.066	.341
Sim.	.642	.097	.125	.398
Ari.	.308	-.053	.551	.448
Voc.	.844	-.080	.064	.281
Com.	.891	-.070	.004	.254
D.S.	-.221	.525	.385	.484
P.C.	.449	.281	-.205	.744
P.A.	.221	.431	.039	.647
B.D.	-.045	.537	.286	.477
O.A.	.128	.786	-.212	.467
Cod.	.088	.129	.586	.469
Maz.	-.009	-.084	.697	.585
%Common Var.	47.85	25.93	26.22	
%Total Var.	45.00	24.39	24.66	

Intercorrelations Among Factors

	I	II
Factor II	.449	
Factor III	.591	.614

Digit Span on both Factors II and III; Arithmetic on both Factors I and III; and Mazes on Factor III. The modifications tested are described below.

Freeing Arithmetic's Factor I loading significantly improved the model ($\chi^2_1=7.22$, $p<.01$) as did allowing Similarities to load on Factor II ($\chi^2_1=5.83$, $p<.02$). These modifications were the result of large partial derivatives for the parameters on previous tests. Unlike the corresponding loading for the total sample, the loading of Similarities on Factor II appears to be based on large correlations with several Performance Scale tests, as determined by examining Table 10.

Picture Completion's loading on the first factor was freed as suggested by the Promax result and the pattern of partial derivatives on earlier tests. The improvement in the model's fit was significant ($\chi^2_1=4.30$, $p<.05$) and the subtest's Factor II loading was fixed to 0 without significant decrement in the model's fit to the data ($\chi^2_1=1.56$, $p>.05$).

Although Mazes' second factor loading was significant, this subtest was allowed to load on Factor III. This test was prompted by the third factor loading for Mazes in the Promax solution. Although the improvement was significant ($\chi^2_1=4.83$, $p<.05$), the resulting second and third factor loadings for Mazes were both large but nonsignificant (-.434±1.3 and .984±1.12, respectively). This result is probably a function of the large correlation between the two factors. Fixing Mazes' second-factor loading resulted in a significant loading on the third factor, without significantly affecting the fit of the model ($\chi^2_1=0.93$, $p<.05$). This model is presented in Table 30.

An examination of the factor pattern in Table 30 reveals a solution in which Kaufman's Verbal Comprehension and Sequencing factors are intact and each include a subtest from the Perceptual-Organization factor. The latter factor has been reduced to three subtests of reasonable size, one of which (Picture Arrangement) is generally conceded to

Table 30

Maximum Likelihood Analysis of WISC-R Subtests
Age Pool 9-10 Years
Final Model for Three Factors:Estimates and Standard Errors[1]

	I	II	III	Ψ
Inf.	.831(.086)	.0	.0	.310(.060)
Sim.	.609(.108)	.222(.108)	.0	.428(.070)
Ari.	.361(.120)	.0	.430(.124)	.494(.082)
Voc.	.849(.085)	.0	.0	.280(.057)
Com.	.843(.085)	.0	.0	.290(.058)
D.S.	.0	.0	.619(.102)	.617(.105)
P.C.	.416(.102)	.0	.0	.827(.122)
P.A.	.0	.600(.104)	.0	.640(.109)
B.D.	.0	.750(.100)	.0	.437(.099)
O.A.	.0	.598(.104)	.0	.642(.109)
Cod.	.0	.0	.742(.099)	.449(.095)
Maz.	.0	.000	.570(.104)	.675(.110)

	Intercorrelations Among Factors	
	I	II
Factor II	.560(.100)	
Factor III	.610(.095)	.847(.080)

[1] Model 21. from Table 26; (χ^2_{48}=51.63, p=.242).

involve verbal sophistication and specific cultural knowledge. Even Bannatyne's more limited Spatial factor is not intact in this solution. The calculation of P.O. or Spatial factor scores for the purpose of comparison to Conceptualization or Sequencing scores is not supported by the results of this analysis.

The extraction and interpretation of a third factor is of dubious value in the light of examination of the intercorrelations among factors. The extremely high correlation between Factors II and III ($\phi_{3,2}$=.847±.160) makes it difficult to argue that these factors are separate constructs.

Summary for Age Pool 9-10 Years

Although principal components analysis suggested that three factors were necessary to account for the common variance, maximum likelihood analysis indicated that two factors were sufficient.

In the two and three-factor solutions, Picture Completion loaded on the first factor, rather than Factor II. Since this subtest is normally considered a strong component of the Perceptual Organization factor and of Bannatyne's Spatial factor, the interpretation of this factor must be reconsidered. This issue is discussed in Chapter V.

The extremely high correlation between Factors II and III undermines the importance of the third factor and the practice of deriving and comparing factor scores based on Kaufman's or Bannatyne's clinical models. A two-factor model appears to have the most validity and this model is presented in Table 28.

Age Pool 11-12 Years

Principal component analysis suggested three factors were necessary to explain the common variance. Maximum likelihood analysis supported this conclusion, as summarized in Table 31. The General factor solution (χ^2_{44}=170.27, $p<.001$) and unrestricted orthogonal two-factor solution

Table 31

Summary of Maximum Likelihood Analysis
Age Pool 11-12 Years

Model Description	χ^2	d.f.	Prob.
1. Null Model	455.52	66	.000
2. General Factor	170.27	54	.000
Two Factors			
3. Orthogonal unrestricted	73.99	43	.002
4. Verb., Perf. Scales, orthogonal	113.24	54	.000
5. Verb., Perf. Scales, oblique	98.96	53	.000
6. Free $\lambda_{6,2}$	87.87	52	.001
7. Free $\lambda_{3,2}$	77.57	51	.010
8. Fix $\lambda_{8,1}$ *	78.15	52	.011
Three Factors			
9. Orthogonal unrestricted	44.03	33	.095
10. Kaufman, orthogonal	128.33	54	.000
11. Kaufman, oblique	76.37	51	.012
12. Free $\lambda_{3,1}$	65.26	50	.072
13. Free $\lambda_{12,3}$	57.09	49	.200
14. Fix $\lambda_{12,2}$ *	57.40	50	.220
15. Free $\psi_{7,2}$	50.66	49	.408

* These factor patterns are presented in subsequent tables.

(χ^2_{43}=73.99, p=.002) were unsatisfactory, while the unrestricted orthogonal three-factor solution provided an acceptable fit to the data (χ^2_{33}=65.26, p=.095). Examination of the clinical two- and three-factor models is described below.

Two Factors

The Verbal-Performance model was rejected for both the orthogonal and oblique solutions, although the latter gives a better fit than the former (χ^2_1=15.28, p<.001). This result is expected, given the insufficiency of two-factors. The Promax result in Table 32 contained some specific departures from the clinical model which warranted further investigation. Digit Span has a strong loading on the second factor, while Arithmetic has moderate loadings on both factors. The partial derivatives for the earlier tests are consistent with these Promax results.

Digit Span was allowed to load on Factor II, resulting in a significant improvement to the model's fit (χ^2_1=11.09, p<.001). Arithmetic's loading on this factor was also freed, resulting in further improvement (χ^2_1=10.30, p<.01). The resulting Factor I loading for Digit The resulting Factor I loading for Digit Span was nonsignificant and was therefore fixed to 0 without significant distortion (χ^2_1=0.58, p>.05). Arithmetic's loadings on both factors remained significant. This modified factor solution is presented in Table 33.

Three Factors

The orthogonal and oblique two-factor clinical models were both rejected, although the latter improved upon the former (χ^2_3=51.96, p<.001). The Promax solution presented in Table 34 suggests two possible modifications: Arithmetic loaded on both of Factors I and III; and Mazes loaded on the third factor. Three modifications were suggested by the patterns of partial derivatives among fixed parameters in the previous tests. The first pattern involved large derivatives for Mazes' third-factor loading and the covariance of its uniqueness component with those of Digit

Table 32

Principal Factor Analysis of WISC-R Subtests
Age Pool 11-12 Years
Promax Solution for Two Factors

	I	II	Ψ
Inf.	.779	-.032	.415
Sim.	.766	.068	.361
Ari.	.470	.337	.521
Voc.	.921	-.056	.194
Com.	.725	-.108	.534
D.S.	.064	.544	.657
P.C.	.026	.542	.693
P.A.	-.062	.630	.634
B.D.	.011	.639	.585
O.A.	-.075	.720	.526
Cod.	-.058	.519	.754
Maz.	.034	.445	.787
%Common Var.	52.64	47.37	
%Total Var.	41.65	37.47	

Intercorrelation Among Factors

	I
Factor II	.455

Span and Coding. This pattern suggested freeing Mazes' Factor III loading. Large partial derivatives for Arithmetic's first-factor loading and its covariance with the uniqueness component of Information suggested freeing its Factor I loading. Picture Completion and Similarities were correlated, resulting in a large partial derivative for $\psi_{7,2}$.

Arithmetic's loading on Factor I was freed first, resulting in significant improvement to the model's fit ($\chi^2_1=11.11$, $p<.001$) and allowing the model to be accepted at

Table 33

Maximum Likelihood Analysis of WISC-R Subtests
Age Pool 11-12 Years
Final Model for Two Factors: Estimates and Standard Errors[1]

	I	II	Ψ
Inf.	.757(.088)	.0	.426(.072)
Sim.	.780(.087)	.0	.391(.068)
Ari.	.433(.099)	.354(.103)	.558(.087)
Voc.	.906(.081)	.0	.179(.055)
Com.	.682(.091)	.0	.535(.084)
D.S.	.0	.539(.103)	.710(.112)
P.C.	.0	.521(.104)	.728(.114)
P.A.	.0	.598(.101)	.642(.106)
B.D.	.0	.649(.100)	.578(.101)
O.A.	.0	.661(.099)	.563(.100)
Cod.	.0	.458(.106)	.790(.120)
Maz.	.0	.440(.106)	.806(.121)

Intercorrelations Among Factors

	I
Factor II	.423(.103)

[1]Model 8. from Table 31; (χ^2_{52}=78.15, p=.011).

the .05 level of significance. The other suggested modifications were examined to determine their importance in describing the data.

Mazes was allowed to load on Factor III, resulting in significant improvement to the model's fit (χ^2_1=9.17, p<.01). The second factor loading became nonsignificant as a result and was fixed to 0 without significantly affecting the model's fit (χ^2_1=0.31, p>.05). The model as modified to this point is presented in Table 35.

The covariance of the uniqueness components for Similarities and Picture Completion was freed, significantly

Table 34

Principal Factor Analysis of WISC-R Subtests
Age Pool 11-12 Years
Promax Solution for Three Factors

	I	II	III	Ψ
Inf.	.780	-.021	-.018	.415
Sim.	.769	.115	-.046	.353
Ari.	.463	.091	.307	.507
Voc.	.924	-.020	-.047	.193
Com.	.724	-.129	.009	.532
D.S.	.045	.011	.658	.534
P.C.	.043	.829	-.256	.447
P.A.	-.060	.558	.149	.609
B.D.	.007	.432	.298	.584
O.A.	-.069	.719	.081	.456
Cod.	-.080	-.092	.729	.565
Maz.	.019	.016	.521	.712
%Common Var.	45.96	28.77	25.25	
%Total Var.	41.52	25.97	22.80	

Intercorrelations Among Factors

	I	II
Factor II	.411	
Factor III	.396	.527

improving the fit of the the model ($\chi^2_1=6.74$, $p<.01$). The psychological interpretation of this relationship is discussed in Chapter V.

The factor pattern presented in Table 35 fits Kaufman's clinical model more closely than the solutions for groups examined thus far, including the total sample. Arithmetic's dual loading on the first and third factor is not inconsistent with the results of the U.S. standardization sample and Kaufman's (1979a) interpretive guidelines account for this finding. Mazes' loading on the third factor is not

Table 35

Maximum Likelihood Analysis of WISC-R Subtests
Age Pool 11-12 Years
Final Model for Three Factors:Estimates and Standard Errors'

	I	II	III	Ψ
Inf.	.758(.088)	.0	.0	.425(.072)
Sim.	.780(.087)	.0	.0	.391(.068)
Ari.	.433(.097)	.0	.417(.104)	.508(.086)
Voc.	.906(.081)	.0	.0	.179(.055)
Com.	.681(.091)	.0	.0	.536(.084)
D.S.	.0	.0	.738(.109)	.456(.119)
P.C.	.0	.591(.103)	.0	.651(.109)
P.A.	.0	.623(.102)	.0	.612(.106)
B.D.	.0	.636(.102)	.0	.596(.105)
O.A.	.0	.739(.099)	.0	.435(.100)
Cod.	.0	.0	.531(.110)	.718(.119)
Maz.	.0	.0	.518(.110)	.731(.120)

Intercorrelations Among Factors

	I	II
Factor II	.389(.107)	
Factor III	.360(.118)	.589(.107)

'Model 14. from Table 31; (χ^2_{50}=57.40, p=.220).

consistent with Kaufman's model and implications of this finding for the interpretation of the subtest and the third factor are discussed in Chapter V. However, the ten subtests normally administered do fall within the hypothesized factors. Bannatyne's Conceptualization, Spatial, and Sequencing factors are intact, although the tests omitted in Bannatyne's model (Picture Arrangement and Mazes) do not load on the factors he suggested (Bannatyne, 1974).

Summary for Age Pool 11-12 Years

Two factors are insufficient to describe the data and the two-factor clinical model is further undermined by the loading of Digit Span with the Performance Scale subtests. The three-factor solution fits Kaufman's model very closely. The loading of Mazes on the third factor is the only clear departure from this model. This departure has implications for the definition of the third factor which will be discussed in Chapter V.

Principal factor and maximum likelihood analysis were also performed on the separate correlation matrices of the 11 and 12 year age groups. This was a precautionary measure prompted by the relatively weaker evidence for the equality of the two covariance matrices (see Table 8). The results for the 11 year age group were similar to the results for the 11-12 year age pool. The results for the 12 year age group were affected by a nonsignificant negative correlation between Picture Completion and Coding, which was reflected in a large partial derivative for the corresponding parameter in Ψ. A good fit (i.e., a nonsignificant χ^2 for the overall model) could not be attained without freeing $\psi_{11,7}$. Although freeing this parameter resulted in a nonsignificant χ^2 and significant loadings in a factor structure similar to that accepted for the 11 year age group and 11-12 year age pool, this modification created linear dependencies in Ψ. The tendency for Ψ to become nonpositive definite in small-sample cases was noted by Lawley and

Maxwell (1971). Diagnosis of the difficulty as a sample size problem was also suggested by the fact that the fit of the model revolved around a nonsignificant correlation. Therefore, the decision to consider the covariance matrices for the two age groups homogenous, as indicated by the results of the tests in Table 8, was accepted.

Age Pool 13-14 Years

Although principal components analysis suggested the extraction and interpretation of four factors, maximum likelihood analysis indicated that two factors was sufficient to explain the common variance. Table 36 summarizes the maximum likelihood analysis. The General factor model was rejected (χ^2_{54}=113.12, $p<.001$) but the unrestricted orthogonal two-factor model provided a marginally satisfactory fit to the data (χ^2_{43}=58.65, $p=.056$). However, three factors provided a significantly better (χ^2_{10}=32.17, $p<.001$) fit to the data than two factors. The examination of the clinical models is described below.

Two Factors

The oblique solution for the Verbal-Performance model provided a significantly better fit than the orthogonal solution (χ^2_{1}=21.01, $p<.001$), although both models were unsatisfactory. The Promax solution presented in Table 37 suggests that Picture Arrangement should be allowed to load on the first factor. This modification was supported by a large partial derivative for the fixed parameter on previous

Table 36

Summary of Maximum Likelihood Analysis
Age Pool 13-14 Years

Model Description	χ^2	d.f.	Prob.
1. Null Model	389.10	66	.000
2. General Factor	113.12	54	.000
Two Factors			
3. Orthogonal unrestricted	58.65	43	.056
4. Verb., Perf. Scales, orthogonal	95.65	54	.000
5. Verb., Perf. Scales, oblique	74.64	53	.027
6. Free $\lambda_{8,1}$	69.12	52	.056
7. Free $\lambda_{7,1}$	66.52	51	.071
8. Fix $\lambda_{7,1}$; Free $\lambda_{6,2}$	66.89	51	.067
9. Fix $\lambda_{8,2}$	70.79	52	.043
10. Fix $\lambda_{6,2}$	72.87	53	.036
11. Free $\lambda_{8,2}$ *	69.08	52	.057
Three Factors			
12. Orthogonal unrestricted	26.48	33	.782
13. Kaufman, orthogonal	103.89	54	.000
14. Kaufman, oblique	61.27	51	.154
15. Free $\lambda_{1,3}$	53.23	50	.351
16. Free $\lambda_{8,1}$ *	47.05	49	.553
17. Fix $\lambda_{8,2}$	51.44	50	.417
18. Free $\lambda_{12,3}$; $\lambda_{8,2}$	46.70	48	.526
19. Fix $\lambda_{12,2}$	55.88	49	.232

* These factor patterns are presented in subsequent tables.

tests. Freeing this loading resulted in a significant improvement to the model ($\chi^2_1=5.52$, $p<.02$), which was now acceptable at the .05 level of significance.

Picture Completion was allowed to load on the first factor as a result of a large partial derivative for this parameter, but this modification did not result in significant improvement ($\chi^2_1=2.60$, $p>.05$) and was not retained. Digit Span was allowed to load on the second

Table 37

Principal Factor Analysis of WISC-R Subtests
Age Pool 13-14 Years
Promax Solution for Two Factors

	I	II	Ψ
Inf.	.755	-.001	.431
Sim.	.727	.030	.451
Ari.	.481	.089	.722
Voc.	.927	-.075	.198
Com.	.836	-.105	.370
D.S.	.247	.182	.865
P.C.	.191	.525	.598
P.A.	.344	.225	.761
B.D.	.048	.787	.344
O.A.	-.087	.604	.674
Cod.	.128	.278	.874
Maz.	-.080	.454	.820
%Common Var.	65.16	34.84	
%Total Var.	49.82	26.64	

Intercorrelation Among Factors

	I
Factor II	.449

factor, but this modification did not affect the model's fit to the data ($\chi^2_1=2.23$, $p>.05$) and was not retained.

Picture Arrangement's second-factor loading was fixed to 0. This modification resulted in a significantly poorer fit to the data ($\chi^2_1=3.90$, $p<.05$) and Picture Arrangement was allowed to load on both factors. The resulting 95% confidence interval for the second-factor loading is .254±.258, and the significance of the above χ^2 test was marginal, since $\chi^2_{1(.05)}=3.84$. Given the present concerns regarding small sample size, the loading was retained and is included in the model. The reader should note that both Picture Arrangement loadings are small and that the subtest's communality is only .767.

The result of the juggling described above (Model 11 in Table 36) was identical to Model 6 in the summary table. Model 11 is presented in Table 38. This model was only marginally significant (p=.057) and yet appears to be the best model attainable with two factors. The largest remaining partial derivatives were for uniqueness covariances between pairs of subtests which load on Kaufman's third factor. This pattern of modification indices suggested the need for a third factor rather than specific modifications to a two-factor model. Examination of the three-factor solution is described below.

Three Factors

Although the orthogonal model based on Kaufman's factors was rejected, the oblique Kaufman model was acceptable without modification (χ^2_{51}=61.27, p=.154). Several modifications were tested as a result of principal factor analysis results and the pattern of partial derivatives for fixed parameters. The results of the Promax rotation of the principal factor solution are presented in Table 39. The most salient departures from the clinical model are: a large third-factor loading for Information; a large loading on Factor I for Picture Arrangement, paired with a lower loading for that subtest on Factor II.

Table 38

Maximum Likelihood Analysis of WISC-R Subtests
Age Pool 13-14 Years

Final Model for Two Factors: Estimates and Standard Errors[1]

	I	II	Ψ
Inf.	.734(.092)	.0	.462(.077)
Sim.	.746(.091)	.0	.444(.075)
Ari.	.456(.102)	.0	.792(.118)
Voc.	.913(.083)	.0	.167(.054)
Com.	.768(.090)	.0	.410(.072)
D.S.	.0	.321(.113)	.897(.134)
P.C.	.0	.607(.107)	.631(.112)
P.A.	.302(.121)	.254(.129)	.767(.116)
B.D.	.0	.808(.103)	.347(.109)
O.A.	.0	.502(.109)	.748(.120)
Cod.	.0	.338(.113)	.886(.133)
Maz.	.0	.381(.112)	.855(.130)

Intercorrelations Among Factors

	I
Factor II	.505(.099)

[1] Model 11. from Table 36; (χ^2_{52}=69.08, p=.057).

A third-factor loading for Information was also suggested by large partial derivatives in Ψ among Information, Arithmetic, and Digit Span in previous LISREL tests. Freeing this loading resulted in a significant improvement (χ^2_1=8.04, p<.01) and the modification was retained.

Arithmetic was allowed to load on the first factor. The improvement in the model was significant (χ^2_1=6.18, p<.05). The model as modified to this point was chosen as the best description of the data matrix. The estimates for Λ, Φ, and

Table 39

Principal Factor Analysis of WISC-R Subtests
Age Sample 13-14 Years
Promax Solution for Three Factors

	I	II	III	Ψ
Inf.	.504	-.070	.416	.393
Sim.	.799	.074	-.098	.389
Ari.	.047	-.056	.717	.472
Voc.	.814	-.086	.188	.197
Com.	.917	-.048	-.129	.294
D.S.	-.109	.053	.600	.678
P.C.	.285	.539	-.077	.556
P.A.	.430	.255	-.102	.719
B.D.	-.037	.721	.243	.341
O.A.	.028	.615	-.106	.640
Cod.	-.081	.194	.374	.814
Maz.	-.109	.420	.107	.820
%Common Var.	48.15	27.47	24.38	
%Total Var.	42.82	24.43	21.68	

Intercorrelations Among Factors

	I	II
Factor II	.376	
Factor III	.524	.307

Ψ are presented in Table 40. Subsequent modifications included: fixing Picture Arrangement's second-factor loading, which resulted in significant distortion and was reversed; and switching Mazes' loading to the third factor, which was also inappropriate. These tests are detailed in Table 36.

Both of the modifications to Kaufman's model have been noted in the literature pertaining to other samples (see Chapter II), but neither fits the clinical model. It could be argued that these factors could be labelled Verbal

Table 40

Maximum Likelihood Analysis of WISC-R Subtests
Age Pool 13-14 Years
Final Model for Three Factors:Estimates and Standard Errors[1]

	I	II	III	Ψ
Inf.	.540(.110)	.0	.332(.118)	.396(.073)
Sim.	.753(.091)	.0	.0	.444(.075)
Ari.	.0	.0	.759(.119)	.512(.139)
Voc.	.915(.083)	.0	.0	.181(.057)
Com.	.775(.090)	.0	.0	.406(.071)
D.S.	.0	.0	.515(.115)	.615(.125)
P.C.	.0	.606(.108)	.0	.689(.114)
P.A.	.307(.117)	.265(.125)	.0	.669(.115)
B.D.	.0	.815(.107)	.0	.451(.119)
O.A.	.0	.518(.110)	.0	.606(.120)
Cod.	.0	.0	.406(.117)	.713(.131)
Maz.	.0	.379(.113)	.0	.779(.131)

Intercorrelations Among Factors

	I	II
Factor II	.461(.105)	
Factor III	.531(.112)	.440(.127)

[1] Model 16. from Table 36; (χ^2_{49}=47.05, p=.553).

Comprehension, Perceptual Organization, and Sequencing with as much confidence as they are so labelled in the United States and southern Canada. This statement is offered in the context of the challenges to these labels described in Chapter II.B.

Four Factors

The Promax solution for four factors is provided in Table 41. Given the difficulty with obtaining a final solution with four factors for the total sample, the only reason for presenting this solution was the suggestion of

Table 41

Principal Factor Analysis of WISC-R Subtests
Age Sample 13-14 Years
Promax Solution for Four Factors

	I	II	III	IV	Ψ
Inf.	.508	.421	-.068	-.032	.393
Sim.	.798	-.078	.137	-.088	.382
Ari.	.040	.740	.052	-.142	.454
Voc.	.822	.188	-.112	-.006	.196
Com.	.923	-.124	-.041	-.036	.294
D.S.	-.099	.576	-.143	.214	.662
P.C.	.296	-.103	.200	.411	.534
P.A.	.441	-.121	.043	.244	.705
B.D.	-.040	.245	.485	.312	.340
O.A.	-.002	-.036	.831	-.195	.492
Cod.	-.090	.394	.230	-.032	.799
Maz.	-.080	.028	-.182	.703	.643
%Common Var.	45.08	22.85	17.23	14.55	
%Total Var.	43.11	21.85	16.76	13.91	

Intercorrelations Among Factors

	I	II	III
Factor II	.527		
Factor III	.365	.277	
Factor IV	.355	.320	.610

four factors in the principal components analysis. The Promax results indicated a splitting of Kaufman's Perceptual Organization factor into two factors. The first of these contained high loadings for Block Design and Object Assembly, while the second factor includes Picture Completion, Block Design, and Mazes. These factors have a large correlation ($\phi_{4,3}$=.610) and the author is unaware of any theoretical or historical basis for predicting such a split. Labels such as "Visual Construction" and "Scanning

Strategies" are suggested by the task demands common within the two sets of subtests, but such labels are purely conjectural.

Attempts to replicate the four factors with maximum likelihood analysis was frustrated by the same difficulties encountered with the tests of the Bannatyne model mentioned above. Convergence to a minimum of F (the loss-ratio function) was not attained after 250 iterations for either the unrestricted orthogonal solution or the direct test of the model suggested by the Promax result. Therefore, the three-factor solution remains as the best description of the correlation matrix for the 13-14 year age pool.

Cross-Validation of Total Sample Results

The best-fitting solution for the total sample, presented in Table 18, was tested for goodness of fit to each of the four age pool correlation matrices. The overall fit of the model to each age pool is described below. New implications are discussed for interpretation of both the total sample findings and those of the age pool in question.

Age Pool 7-8 Years

The estimates for Λ, Φ, and Ψ are presented in Table 42. Comparisons between this solution and the factor pattern presented in Table 25 reveal few substantial differences. Model-discrepant loadings which were small for the total sample were nonsignificant for this age pool. These include $\lambda_{7,1}$, $\lambda_{2,2}$, and $\lambda_{1,3}$. Digit Span had nonsignificant loadings

Table 42

Cross-validation of ML Results for Total Sample
Age Pool 7-8 Years
Three Factors: Estimates and Standard Errors

	I	II	III	Ψ
Inf.	.599(.254)	.0	.195(.255)	.401(.077)
Sim.	.626(.114)	.162(.114)	.0	.516(.094)
Ari.	-.023(.406)	.0	.808(.404)	.379(.113)
Voc.	.931(.094)	.0	.0	.133(.057)
Com.	.785(.103)	.0	.0	.384(.077)
D.S.	.0	.0	.639(.117)	.592(.117)
P.C.	.129(.124)	.669(.133)	.0	.479(.118)
P.A.	.434(.113)	.465(.117)	.0	.463(.093)
B.D.	.0	.722(.123)	.0	.479(.127)
O.A.	.0	.582(.126)	.0	.661(.134)
Cod.	.0	.0	.355(.125)	.874(.151)
Maz.	.0	.0	.519(.120)	.730(.132)

Intercorrelations Among Factors

	I	II
Factor II	.329(.150)	
Factor III	.868(.091)	.548(.146)

$\chi^2_{46}=48.13$, p=.387

on both the first and third factor in the cross-validation result. Mazes loaded on the third factor. The last two results are less surprising when one considers the high correlation among Factors and the fact that both PC and ML analysis indicatd that only two factors were required to explain the common variance.

Although the overall model had an acceptable χ^2 value, the pattern of nonsignificant loadings supports the original conclusion that interpretation of three factors is misleading.

Age Pool 9-10 Years

The result for the cross-validation with this age group is presented in Table 43. As with the 7-8 year age pool, loadings which were not included in the modifications tested earlier for this age pool were generally nonsignificant, including $\lambda_{8,1}$, $\lambda_{7,2}$, and $\lambda_{1,3}$. Picture Completion's first-factor loading remained significant.

The only difference between the cross-validation and exploratory results was Mazes' loading on Factor III on the latter test. Given the ML indication that only two factors should be interpreted, and given the high correlation among the factors, this result could be expected. Mazes did not have a stable third factor loading for either of the two younger age pools. It appeared to have large correlations with each of the factors and can therefore have a significant loading on any single factor. The interpretation of this subtest for children age 7 years to 10 years is not discernable from the factor analytic results.

Age Pool 11-12 Years

The LISREL estimates for the cross-validation test on the 11-12 year age pool are provided in Table 44. These results correspond exactly to the final result from the ML exploratory analysis for this age pool. All loadings which were significant for the latter test were significant in the cross-validation test. All other loadings were nonsignificant. The size and ranking of factor correlations and the size of the uniqueness coefficients are comparable

Table 43

Cross-validation of ML Results for Total Sample
Age Pool 9-10 Years
Three Factors:Estimates and Standard Errors

	I	II	III	Ψ
Inf.	.737(.105)	.0	.138(.101)	.321(.059)
Sim.	.611(.090)	.254(.099)	.0	.425(.070)
Ari.	.349(.115)	.0	.453(.120)	.493(.081)
Voc.	.856(.085)	.0	.0	.268(.058)
Com.	.851(.085)	.0	.0	.275(.058)
D.S.	.0	.0	.626(.102)	.608(.104)
P.C.	.315(.119)	.197(.125)	.0	.807(.119)
P.A.	.214(.114)	.459(.121)	.0	.656(.105)
B.D.	.0	.791(.102)	.0	.374(.106)
O.A.	.0	.601(.105)	.0	.638(.110)
Cod.	.0	.0	.734(.098)	.461(.095)
Maz.	.0	.0	.568(.104)	.678(.110)

	Intercorrelations Among Factors	
	I	II
Factor II	.442(.116)	
Factor III	.570(.102)	.823(.083)

$\chi^2_{46}=44.83$, $p=.521$

with the "best-fitting" three-factor model for this sample.

The smaller loadings from the total sample results were nonsignificant, leaving a model which conforms more closely to Kaufman's clinical model than does the total sample result. The third-factor loading for Mazes was significant, as it was in the results of the exploratory analysis for this age pool.

Age Pool 13-14 Years

The cross-validation results for this age pool, presented in Table 45, correspond closely to those of the

Table 44

Cross-validation of ML Results for Total Sample
Age Pool 11-12 Years
Three Factors:Estimates and Standard Errors

	I	II	III	Ψ
Inf.	.743(.095)	.0	.031(.094)	.432(.073)
Sim.	.728(.092)	.127(.087)	.0	.388(.067)
Ari.	.424(.096)	.0	.435(.104)	.506(.086)
Voc.	.912(.081)	.0	.0	.168(.057)
Com.	.685(.091)	.0	.0	.531(.084)
D.S.	.0	.0	.736(.108)	.459(.118)
P.C.	.032(.108)	.585(.115)	.0	.643(.108)
P.A.	-.027(.108)	.630(.116)	.0	.614(.109)
B.D.	.0	.646(.102)	.0	.583(.105)
O.A.	.0	.722(.100)	.0	.479(.101)
Cod.	.0	[illegible]	.528(.109)	.721(.119)
Maz.	.0	.0	.517(.110)	.732(.120)

Intercorrelations Among Factors

	I	II
Factor II	.361(.121)	
Factor III	.339(.121)	.604(.106)

$\chi^2_{46}=55.18$, p=.166

exploratory analysis. Information's third-factor loading and Picture Arrangement's first-factor loading were significant in both analyses. Although Picture Completion did not load on Factor I in the exploratory analysis, this loading was marginally significant in the cross-validation result, with a 95% confidence interval of .227±.222. Other departures from Kaufman's model did not generalize from the total sample to the 13-14 year sample. Arithmetic and Picture Completion did not load on the first factor and Similarities' second-factor loading was also nonsignificant.

Table 45

Cross-validation of ML Results for Total Sample
Age Pool 13-14 Years
Three Factors:Estimates and Standard Errors

	I	II	III	Ψ
Inf.	.575(.109)	.0	.323(.111)	.403(.074)
Sim.	.711(.098)	.102(.092)	.0	.429(.073)
Ari.	.155(.145)	.0	.592(.154)	.545(.126)
Voc.	.907(.084)	.0	.0	.178(.058)
Com.	.786(.090)	.0	.0	.381(.070)
D.S.	.0	.0	.560(.124)	.686(.134)
P.C.	.227(.111)	.438(.121)	.0	.682(.110)
P.A.	.345(.112)	.228(.116)	.0	.770(.115)
B.D.	.0	.880(.124)	.0	.225(.169)
O.A.	.0	.533(.112)	.0	.716(.122)
Cod.	.0	.0	.431(.122)	.814(.134)
Maz.	.0	.0	.227(.124)	.949(.141)

Intercorrelations Among Factors

	I	II
Factor II	.375(.114)	
Factor III	.435(.156)	.489(.131)

χ^2_{46}=49.96, p=.319

Mazes' third factor loading was not significant. As the reader may note from Table 40, Mazes loaded on the second factor for this age pool, as predicted from Kaufman's model.

Picture Arrangement's second-factor loading was nonsignificant in the cross-validation result, while it was only marginally significant in the best-fitting result from the exploratory analysis. The addition of free parameters into the model could easily have affected the estimates to this small degree, although these parameters are nonsignificant.

Summary of Cross-validation Results

The main contribution of this portion of the analysis may have been confirmation of the models chosen as "best-fitting" for each age pool. Parameters which were significant in the total sample analysis were generally significant in the cross-validation for a given age pool only if those parameters were free and significant in that age pool's best-fitting model. Mazes were an exception to this rule. The redundancy of the third factor for the two youngest age pools was reflected in the instability of the results for Mazes.

It may be argued that the discrepancies between the results for the total sample and the cross-validations is a function of the difference in sample size and resulting statistical power. Comparison of Table 14 with the factor patterns for the cross-validation tables reveals that while the estimates for loadings were often comparable across samples the standard errors of estimates were much larger for the age pools than for the total sample. For example, $\lambda_{3,1,7-8} = .195$ and $\lambda_{3,1,ts} = .150$. The standard errors for these parameters are .255 and .056 respectively. The limitations on the validity of the earlier confirmatory and exploratory analyses with the age pool correlation matrices as a result of sample size extend to the cross-validation analyses. Some reassurance may be gained, however, from the fact that each of the small loadings in the total sample's three-factor pattern matrix which deviated from Kaufman's model are replicated by at least one age pool. Age pools 9-10 and 11-12 had significant loadings for $\lambda_{3,1}$; age pools 9-10 and 13-14, for $\lambda_{7,1}$; age pools 7-8 and 13-14, for $\lambda_{8,1}$; age pool 9-10, for $\lambda_{2,2}$; and age pool 13-14, for $\lambda_{1,3}$. It is the author's conclusion that while the total sample's larger numbers will allow a more powerful test of the various models, the age pool analyses are capable of detecting relatively subtle deviations from a hypothesized model. A failure to replicate the acceptance or rejection of a particular loading cannot be summarily dismissed on the basis of the sample size for the age pool.

Summary of Factor Analytic Results

The number of factors required to reproduce the various correlation matrices varied across sample and method of factoring. Three factors were required for the pooled total sample matrix, according to both maximum likelihood and principal components analysis. The number required for each age pool is summarized in Table 46 and discussed below.

Modifications to the clinical models were guided by the modification indices provided by **LISREL**, particularly the partial derivatives for fixed parameters. Other guides included the principal factor analysis results for the given age pool and trends among the other age pools. The model indicated by maximum likelihood as the best factor model for that age pool was typically very similar to that sample's principal factor analysis result when convential criteria

Table 46

Number of Factors Suggested by PC and ML Methods

Method	Sample Total	 7-8	 9-10	 11-12	 13-14
Principal Components	3	2	3	3	4
Maximum Likelihood	3	2	2	3	2

for identifying salient factors (such as loadings ≥ .30) were employed.

The null model of Bentler and Bonett (1980) was rejected for every sample. The fact that all clinical models required modification for almost all of the samples demonstrates that the sample size was not so small as to preclude rejection of a clearly inappropriate model. The results of the null model are somewhat redundant under these circumstances.

The single-factor general intelligence model was rejected in every sample, indicating that one factor was insufficient to explain the data matrix.

Orthogonality of the restricted factor solution was rejected in every case, regardless of the number of factors involved.

As stated above, three factors were required for the total sample's pooled matrix and Kaufman's model required modifications. The clearest departure from Kaufman's model was Mazes' loading on the third factor. Arithmetic, Picture Completion, and Picture Arrangement had loadings on Factor I; Similarities on Factor II; and Information on

Factor III. Although two factors were insufficient for this sample, it should be noted that Digit Span loaded on the Performance factor when the Verbal-Performance model was tested.

The best fit for the 7-8 year age pool was a two-factor model in which Mazes loaded on the Verbal factor and Picture Arrangement loaded on both factors. A third factor was not stable as it was highly correlated with the other factors.

The 9-10 year age pool had a poor fit to both the two- and three-factor clinical models. Picture Completion loaded with the Verbal subtests on both models while Digit Span and Arithmetic loaded with the Performance subtests on the two-factor model. While the Verbal Comprehension and Sequencing factors were intact on the three-factor model, the extremely high correlations among the factors made interpretation of the third factor misleading. The VC loading for Picture Completion undermines the integrity of the Perceptual Organization factor.

Three factors are required for both the 11-12 and 13-14 year age pools. Kaufman's model fit the data much more closely than for the younger age pools. Mazes loads on the third factor for the 11-12 pool, while Picture Arrangement loads on Factor I for the 13-14 pool.

There are three interesting trends regarding individual subtests.

1. Picture Completion and Similarities shared a relatively large correlation which seemed to effect factor loadings

for these tests.

2. Digit Span was inconsistent in regard to factor loadings across age groups. It loaded on the Verbal factor for age pool 7-8, but on the Performance factor for the other age pools' two-factor solutions. It loaded on the third factor, as hypothesized, for three-factor solutions but these solutions were not meaningful for the two younger age pools.
3. Whereas Digit Span remained in the Performance factor for all three of the older age pools, the loadings for Picture Arrangement and Mazes fluctuated with increasing age in both two- and three-factor solutions. Picture Arrangement loaded on both the Verbal and Performance factors at age 7-8 years, but loaded only on the Performance factor at age 9-10 years. At age 11-12 years, this subtest loaded only on PO in the three-factor solution, whereas its predominant loading at age 13-14 years was on the VC factor. Mazes did not have stable loadings for three-factor solutions for the 7-8 and 9-10 age pool. The subtest loaded on the Sequencing factor in age pool 13-14, but loaded on the Spatial factor for age pool 11-12. It loaded on the Verbal factor for age pool 7-8, but on the Performance factor for age pool 9-10. In short, these subtests did not have the consistency required for confident interpretation within a factor.

The lack of a stable Performance factor across the first two age pools strongly suggests a lack of construct validity for the WISC-R at ages 7 to 10 years. The Kaufman three-factor model, with minor modifications, appears to have validity for Arctic children aged 11 to 14 years, although Mazes may not be interpreted as a Perceptual Organization subtest. Bannatyne's Acquired Knowledge factor has no validity as a construct distinct from Conceptualization. The psychological implications of these conclusions are discussed in greater detail in Chapter V. The remainder of Chapter IV describes an additional analysis which was conducted on Digit Span Forward and Backward to explore the unexpected tendency for the test to load on a spatial factor.

C. Digit Span Analysis

Extended analysis was conducted on Digit Span to address concerns relating to the two-part structure of the test. Literature regarding these concerns is reviewed in Section B of Chapter II. If Digit Span Forward (DSF) and Backward (DSB) measure verbal and visual memory skills, respectively, these two components of the subtest could be expected to load on the Verbal and Performance factors, respectively, in the two-factor solution. The common memory or sequencing factor underlying DSF and DSB could result in the two components loading together in the third factor of the three factor solution, overriding differences on the verbal-visual dimension. Digit Span loaded on the Performance factor of the two-factor solution for the total sample and the three oldest age pools. It loaded on the third factor of the three-factor solution for all samples. The factors underlying the two parts of the test were explored by factor analyzing DSF and DSB with the remaining 11 WISC-R subtests.

Scaled scores were not calculated for the separate parts of D.S. in the Mulcahy and Watters (1982) norming study, so raw scores on DSF and DSB were correlated with the remaining subtests for each of the eight age groups. These correlation matrices were factored by principal factor analysis, followed by Varimax and Promax rotations on the factor matrices. This analysis is part of a series of analyses on the Digit Span subtest which is scheduled for publication at a later date. The most salient results for each age group are briefly described below. References to factor names indicate factors resembling those defined by Kaufman, Bannatyne, or the Verbal and Performance Scales. This allows concise descriptions of the results and is not intended to imply either exact fit to those factor structures or acceptance of the cognitive theories associated with them.

The correlations of DSF and DSB with the remaining 11 subtests are provided for each age group in Table 47. The intercorrelations among the latter subtests are available in Mulcahy and Watters (1982). The correlations between DSF and

Table 47

Subtest Correlations with Digit Span Forward and Backward By Age Group

Test	Age Group 7	8	9	10	11	12	13	14
Digit Span Forward								
DSB	.470	.302	.508	.389	.275	.181	.483	.409
Inf.	.195	.331	.432	.265	.280	.210	.446	.175
Sim.	.254	.426	.380	.178	.349	.311	.223	.173
Ari.	.259	.528	.234	.304	.488	.441	.343	.219
Voc.	.308	.488	.260	.133	.285	.207	.293	.241
Com.	.177	.428	.208	.110	.175	.099	.294	.147
P.C.	.124	.325	.183	.091	-.047	.291	.390	-.005
P.A.	.282	.422	.431	.039	.276	.114	.119	.220
B.D.	.202	.103	.481	.247	.139	.284	.451	.117
O.A.	.323	.238	.461	.299	.212	.039	.395	-.081
Cod.	.040	.047	.395	.403	.303	.302	.244	.091
Maz.	-.140	.230	.252	.254	.154	.361	.261	.141
Digit Span Backward								
Inf.	.418	.334	.294	.045	-.042	.117	.324	.145
Sim.	.400	.216	.231	.125	.152	.011	.122	.057
Ari.	.492	.371	.409	.390	.270	.240	.519	.252
Voc.	.595	.390	.164	.133	.174	.060	.191	.080
Com.	.460	.198	.099	.084	-.036	.113	-.090	.053
P.C.	.102	.361	.128	.102	.170	.331	.227	-.044
P.A.	.378	.333	.304	.505	.127	.488	.133	-.112
B.D.	.227	.274	.376	.479	.238	.336	.342	-.042
O.A.	.485	.364	.501	.313	.107	.326	.164	-.473
Cod	.116	.345	.292	.415	.331	.220	.180	.002
Maz.	.449	.188	.255	.368	.244	.517	.151	.003
Critical Values of r for Age Group Ns[1]								
N	34	37	48	49	52	50	53	43
$r_{(.05)}$	.340	.325	.285	.282	.273	.279	.271	.301
$r_{(.01)}$	.437	.418	.368	.365	.354	.361	.351	.389

[1] Adapted from Table I of Glass & Stanley, 1970, p. 536.

DSB are provided in the first half of Table 47. The critical values of the correlation coefficient for a two-tailed test of significance at the .05 and .01 levels are provided at the bottom of the table for the sample size of each age group.

Age 7 Years

Although DSF and DSB had a correlation of .470, DSF was not significantly correlated with any other subtest. DSB had significant correlations with all but P.C. and B.D. and loaded on a Verbal factor in both the two- and three-factor solutions. DSB also loaded on a third factor containing large loadings for only DSF and DSB. DSF did not load on any other factor for either the two- or three-factor solution.

Age 8 Years

The results for this age group conformed more closely to the verbal-visual memory hypothesis. DSF correlated with the Verbal Scale subtests plus Picture Completion and Picture Arrangement and loaded on a Verbal factor, whether two or three factors were rotated. DSB correlated with a number of subtests from both scales and loaded on a Spatial factor in both solutions. The nonsignificant correlation between DSF and DSB invalidates the interpretation of Digit Span as a measure of a single construct.

Age 9 Years

The combination of DSF and DSB received more empirical support from the results of this age group, with the two components significantly intercorrelated and both loading on a Spatial factor in each of the two- and three-factor solutions.

Age 10 Years

Both DSF and DSB loaded on a factor comprised of subtests from Bannatyne's Spatial and Sequencing factors in the two-factor solution and both loaded on a Spatial factor in the three-factor solution, although DSB was significantly correlated with more individual subtests than was DSF.

Age 11 Years

Both DSF and DSB loaded on a Sequencing factor in the three factor solution and their intercorrelation was significant, if unimpressive, at .275. DSF had no large loadings in the two-factor solution, where DSB loaded on a factor including both Spatial and Sequential subtests.

12 Years

Although DSF and DSB were not significantly correlated, each was correlated with a number of Performance Scale subtests and loaded on a Spatial factor in the two- and three-factor solutions.

Age 13 Years

DSF and DSB were correlated at .483 and both loaded on a factor combining Spatial and Sequential tests in the two-factor solution. Both components loaded on a Sequencing factor when three factors were rotated.

Age 14 Years

The results for this age group contradicted those of the 13 year old sample with which they were pooled and theoretical interpretations of the Digit Span subtest and its two components. Although DSF and DSB correlated significantly with each other, DSF did not correlate with any other subtest and DSB had only one significant, but negative, correlation with Object Assembly. DSF and DSB loaded only on a factor for which the only other salient loading was a negative loading for Object Assembly. The two subtest components appeared to share some common variance, but the validity of their contribution to factor scores is threatened by these results. The stability of the third factor for the 14 year age group is therefore threatened as well.

Given the decision to not reject the hypothesis of equality of correlation matrices, summarized in Table 9, the results for the 13-14 year age pool are presumed to generalize to the 14 year group. Principal factor and maximum likelihood analysis performed separately for the two groups indicated that the factor pattern in Table 40 may be generalized to the 13 year age group with more confidence than is justified with the 14 year age group. The severity of this discrepancy must be weighed against the low reliability of factor analytic results with samples of 43 subjects. The difficulty experienced in seeking a proper solution for the 12 year age group, before it was pooled with the 11 year sample, reinforces this caution. Principal factor analysis on the twelve subtests for each of the eight age groups suggested that the factor pattern for the other age pools were closely matched by the results of their member age groups.

Further research on the psychometric properties of the test is currently being considered. This research may examine age trends in score distribution which could affect the correlations among subtests. For the purposes of this study, the three-factor solution identified as the best-fitting model for the 13-14 year age pool is offered as the best interpretation of the WISC-R for children of both age groups. The reader should exercise more caution in applying this solution to the older group, pending the results of the research planned.

Summary of Digit Span Analysis

The analysis of DSF and DSB as separate tests did not support the theory that the former measures verbal memory while the latter measures visual memory. This split was manifested only in the 8 year age group, where DSF loaded with Verbal Scale subtests while DSB loads with Performance Scale subtests and the two Digit Span components were not significantly intercorrelated. In the 7 year age sample, DSB loaded on the Verbal factor, while DSF had no salient loadings and was not significantly correlated with anything but DSB. In all other age groups the two components loaded together on a Spatial factor, a Sequential factor, or a factor comprised of both Spatial and Sequential subtests, depending on the number of factors rotated and whether the results of the maximum likelihood analysis had indicated a two- or three-factor solution.

Although the verbal-visual memory hypothesis was not supported, the low and occasionally nonsignificant correlations between DSF and DSB suggest that Digit Span is not an internally consistent subtest. While they tended to load on the same factor for most age groups, DSF and DSB cannot be considered parallel forms. It is the author's hypothesis that Digit Span measures short-term memory for older children who have no difficulty understanding the directions for the subtest, and that the "visual" strategy of memory rehearsal associated with DSB may be employed with both DSF and DSB. The Verbal factor loadings obtained at ages 7 years and 8 years are hypothesized to be a function of the child's fluency in English and ability to understand the instructions for the subtest. These hypotheses are discussed in greater detail in Chapter V, integrating the above results with the literature cited in Chapter II and clinical observations of the testers.

V. Discussion

Discussion of the results for the present study begins with the presentation of four recommendations related to the hypotheses listed in Chapter II.E. A detailed examination of the unexpected results for some specific subtests is provided to suggest possible explanations of failures to replicate the clinical factor models with the data for the younger age pools. Limitations of the present study are then discussed, followed by suggestions for future research on assessment of the NWT.

The present author has repeatedly noted the difficulties associated with attempting to identify constructs solely on the basis of examination of the matrix of factor loadings. The use of the factor names suggested by Wechsler, Kaufman or Bannatyne is a matter of convenience and does not indicate acceptance of their interpretation of a factor. Discussions of alternative interpretations include ideas on possible empirical tests of those interpretations.

A. General Recommendations

The first recommendation is based on the finding that the scaled scores for several subtests were not normally distributed in various age groups. The second recommendation is based on the results of tests of the homogeneity of covariance matrices within the age pools and across all age groups. The third recommendation is based on the rejection of all of the clinical models for the two youngest age pools, while the final two recommendations concern the limits of interpretability of the evidence for the construct validity of a three-factor model for the two oldest age pools.

1. The validity of interpretation of individual subtests is undermined in several cases by subtest distributions which are significantly nonnormal. Comparisons among individual subtests are invalidated where one or more of the subtests has a nonnormal distribution.

The subtests in question were identified in Chapter IV.A. The effect of the presence of nonnormal distributions on subtest interpretation and comparison is best demonstrated by example. Similarities was identified as having a significantly nonnormal distribution for both the 7 and 8 year age groups. The scaled mean in each case was close to 10 in each case (10.27 and 9.92, respectively). However, half of the 7 year sample obtained a scaled score of 8 on that subtest (corresponding to a raw score of 0). A child who obtained a scaled score of 10 would have scored on the 74%ile at age 7 years; the 81%ile at age 8 years. If the distributions were normal, a score of 10 would have corresponded to the 50%ile at every age. While a score of 10 represents average ability, an 8 year old child who obtains that score has achieved in the top 19% of his age group.

The inferential errors in subtest interpretation which would occurr by assuming a normal distribution would be compounded in the comparison of subtests. Block Design has an almost rectangular distribution at age 8 years. A child who received subtest scores of 10 and 13 on Similarities and Block Design, respectively, would normally be considered to have a relative strength on the latter subtest, as a difference of three scaled score points is considered significant. However, for an 8 year old child in the NWT sample, the corresponding %iles would be 81 and 78, respectively. There is no relative strength and weakness in terms of the child's rank on the test. Since inferences about rank are all that such scaled scores normally allow, the assortment of skewed, flat, and peaked distributions which occur in the sample invalidates individual subtest interpretations and comparisons.

2. The results of the individual age pools should be interpreted separately and these results should take precedence over the results for the total sample.

The failure to reject the hypothesis of homogeneity of all covariance matrices (Hypothesis 2 in Chapter II.E) is evidence in support of the interpretation of the results for the total sample. However, Jöreskog (Note 4) has stated that the test for homogeneity of covariance which is available in LISREL is not sufficiently robust. He supported the interpretation of the separate age pool results for the present study. The results of the cross-validation analyses also indicated that the independently-derived solutions were more reliable factor patterns for the respective age pools. Although the χ^2 values obtained by fitting the total sample a

solution to the age pool correlation matrices were not significantly large, the resulting factor patterns were often less interpretable. For example, allowing a third factor loading on Arithmetic for the 7-8 year age pool resulted in the absence of any significant loadings for that subtest. Even its third factor loading of .808 was nonsignificant. Loadings which were nonsignificant in the age pool results were nonsignificant in the cross-validation results, regardless of their salience in the total sample results. Therefore, the results of analyses on the individual age pools is given more credence in this discussion than the results of the total sample analysis.

3. Construct validity of the WISC-R was not demonstrated for children younger than 11 years of age. Administration and interpretation of the test is not recommended for these children.

As noted in the summary of factor analytic findings in Chapter IV.B, the lack of a stable Performance, Perceptual Organization, or Spatial factor invalidated the test for children aged 7 years to 10 years.

In the 7-8 year age pool, the large loadings obtained for Picture Arrangement and Mazes on a factor with Verbal Scale subtests were deviations from the Verbal-Performance and Kaufman models. However, since Picture Arrangement also loaded on a factor with the other Performance Scale tests, a Performance Scale score could include contributions from the five mandatory subtests.

The results for the 9-10 year age pool are more discrepant from the hypothesized models. Picture Completion's loading with Verbal Scale subtests invalidates the calculation of PO, Spatial, or Performance Scale scores. The Performance factor loadings of Arithmetic and Digit Span in the two-factor solution are also discrepant with the Verbal-Performance model. The final three-factor solution was similar to Kaufman's model. However, the failure to replicate the PO or Spatial factor, the high correlations among the factors, plus the fact that three factors did not significantly improve upon the fit of a two-factor solution (as measured by the comparison of χ^2 values for the respective unrestricted solutions) led to the rejection of the Kaufman model for the 9-10 year age pool.

4. The validity of Kaufman's three-factor model was supported for children in the 11-12 year and 13-14 year age pools. Scores may be calculated on the modified VC, PO, and FD factors or on Bannatyne's Conceptualization, Spatial, and Sequencing factors.

The only modification required for acceptance of the model at 11-12 years was a salient VC loading for Arithmetic, in addition to its FD loading. As noted in Chapter II.B and Chapter II.D, Arithmetic's loadings on these two factors is not inconsistent with Kaufman's (1975) results for the U.S. standardization sample or with his recommendations for interpretation of the WISC-R (Kaufman, 1979a, 1980).

Kaufman's model was accepted without modification for the 13-14 year age pool, although the fit of the model was improved by modifications such as a VC loading for Picture Arrangement and a FD loading for Information. These loadings weaken support for the model, although they are consistent with findings from studies reviewed in Chapter II.B.

Bannatyne's (1974) Conceptualization, Spatial, and Sequencing factors were replicated within the modified three-factor models for each of the two oldest age pools. Calculation of factor scores on the basis of those categories may be more valid than the use of the modified Kaufman factors. This practice would exclude Information, Picture Arrangement, and Mazes from factor interpretations. These three tests account for almost all of the modifications made to the Kaufman factors for the two oldest age groups. The validity of comparison of the Conceptualization and Sequencing factors would be limited by the fact that Arithmetic loaded on both factors at 12-13 years.

5. The clinical utility of the WISC-R for psychoeducational assessment of children in the Keewatin and Kitikmeot regions is necessarily very restricted.

Since the clinical factor models were not supported for children younger than 11 years of age, the test cannot be used as a screening device for early detection of mental retardation or cognitive strengths and weaknesses. The rejection of a single-factor model for each age group undermines the use of FSIQs based on an equal weighting of subtests. Use of subtest scores, factor scores or IQs for simple prediction of school achievement has no empirical support at present. Long-range prediction of the school achievement of bilingual children from psychological test scores was discredited by Cummins (1982) and sharply criticized by Jensen (1980). Although the Kaufman factor model and Bannatyne's original three factors have construct validity support for children aged 11 to 14 years, the clinical and educational utility of scores on those factors remains to be demonstrated.

B. Age Trends in Number and Definition of Factors

The two youngest age pools were clearly differentiated from the oldest pools on the basis of the number of factors chosen for the final maximum likelihood solution. Two factors were chosen for the 7-8 year and 9-10 year pools, while three factor solutions provided the best fit for the older pools. Examination of Table 45 in Chapter IV reveals that the pattern of the number of principal components with eigenvalues greater than 1.0 indicates a trend toward an increase in the number of significant factors with increasing age. Although an unrestricted two-factor MLFA solution was not rejected for the 13-14 year age pool, a three-factor unrestricted solution was significantly better. The correlations among factors in the three-factor solutions decreased with age, suggesting greater independence of the factors. The ranges for these correlations (taken from the final three-factor MLFA solutions for each age pool as tabled in Chapter IV) were .427 to .853., .560 to .847, .389 to .589, and .440 to .531, for the 7-8 year, 9-10 year, 11-12 year, and 13-14 year age pools, respectively. It seems safe to conclude that the number of factors required to explain the correlations among WISC-R subtests increases from two to three over the 7 years to 14 years age span.

The increase in the number of factors underlying WISC-R subtests appears to support Garrett's (1946) differentiation hypothesis, i.e, that intellectual abilities become increasingly specialized throughout childhood and adolescence. Comparison of the factor solutions for the four age pools does not provide a direct test of the differentiation hypothesis, which would require longitudinal comparisons. However, the cross-sectional comparisons available from the NWT results are consistent with the hypothesis.

As stated in Chapter II.A, Garrett believed that the general factor which gradually declined in importance reflected linguistic ability. There were no general factor solutions accepted in the present study. However, there were several occurrences of supposedly nonverbal subtests, such as Picture Completion or Mazes, loading with the Verbal Scale subtests. While such occurrences appear to support the theory of a linguistic general factor, there is evidence from the pattern of results and from observations of testing behavior that this factor reflected familiarity with English rather than the ability to use language per se. This evidence is presented in the discussion of patterns for certain individual subtests, which is presented in Chapter V.C.

The most puzzling aspect of the age trends in closeness of fit to the clinical models is the finding that the Verbal-Performance model was more accurate for the 7-8 years age pool than for the 9-10 year age pool. This finding appears to contradict a language-familiarity explanation for the generally poor fit of the models for these two age pools. The older children would be expected to have had more school experience and therefore more exposure to English. However, given the level of absenteeism in NWT schools (Watters, 1980, in press), the correlation between age and days spent in school is probably far from perfect. The patterns of rates of absenteeism for various age groups, cohorts, and villages could be gathered from the records of the NWT schools and Department of Education. Even if it could be determined that children in the 9-10 year age pool (or subsets of that sample) had been absent so frequently that their exposure to English in the school was not more extensive than that of children in the 7-8 year age pool, this finding would not prove that variance in English fluency was associated with discrepancies from the clinical factor models.

Another possible source of disturbances in the the factor models may have been the switch from Inuktitut to English classroom instruction. The procedures followed in introducing English instruction varies widely across the villages, as was described in Chapter III.A. Schools in the Keewatin region offered Inuktitut instruction for 50% of the curriculum in the first three grades, while Kitikmeot schools tended to offer all the curriculum in English, with Inuktitut offered as a language course. Many children aged 9 to 10 years and attending Keewatin schools would have been undergoing an important shift in the amount of instruction taught in their first language. Children of the same age in Kitikmeot schools would have experienced that shift earlier in their education. As discussed in Chapter II.C, the effectiveness of immersion instruction in a second language appears to be related to numerous factors affecting opportunities for children to use their first language outside the school, thereby allowing continuity of linguistic ability (Cummins, 1978, 1979; Toukamaa & Skutnabb-Kangas, 1977). The interaction of factors such as a decrease in Inuktitut instruction, variance in the amount of English instruction received at the time of WISC-R testing, and opportunities for developing skills in each language outside of school may have affected the subtest correlations for the 9-10 year age pool. By combining the school records of subjects in this age pool with information on the procedures for beginning English instruction in each school, the age at which most subjects began English instruction could be determined. However, data on factors such as the amount of English spoken outside the classroom are not available from the data gathered in the Mulcahy and Watters (1982) norming project. The relationship of WISC-R factor definitions to language of instruction must remain speculative at this point.

Another environmental factor which may have affected the age trends in factor definition is exposure to television. Television was introduced to many arctic communities in the late 1970's and early 1980's. As noted in Chapter II.C, the availability of television was associated with improved school readiness skills on the part of children and increased knowledge of, and interest in, national and international events on the part of adults (O'Connell, 1975; Watson, 1980). Children in the 9-10 year age pool would have had less opportunity for preschool exposure than children in the 7-8 year age pool. Variance in the availability of television across age pools and across villages, particularly in regard to daytime programs such as Sesame Street, may have been associated with the covariance among WISC-R scores for the various age pools. A test of this hypothetical explanation would require a quasi-experimental study in which various cohorts were tested on several occasions before and after the introduction of television to the community. Comparisons to a valid control group, which had not yet received television, would be required. Even then, quasi-experimental studies typically involve more threats to internal validity than true experiments. The fact that most communities in the Keewatin and Kitikmeot regions now have television, plus the impracticality of numerous WISC-R testings, preclude administration of such an experiment.

Failure to replicate the clinical factor models with the subtest correlation matrices for the two youngest age pools led to the conclusion that the WISC-R did not have evidence of construct validity for clinical use with children younger than 11 years of age. Possible sources of disturbances in the factor models were discussed above, but these explanations were highly speculative and offered only as hypotheses. Further discussion of trends on the factor models focuses on some unexpected loadings and correlations obtained for specific subtests.

C. Individual Subtest Anomolies

In the summary to the factor analytic results for the present study, three sets of unexpected results were noted involving certain individual subtests. These results included the strong residual relationship between Similarities and Picture Completion, the internal consistency and factor definition of Digit Span, and the difficulty encountered in identifying stable factor loadings for Picture Arrangement and Mazes.

Similarities and Picture Completion

There was a tendency for Similarities and Picture Completion to have large residual correlations, i.e, to have covariances which were not adequately accounted for by the correlation between their respective factors. This tendency was reflected in large partial derivatives for the covariances of the error terms for these tests, $\psi_{7,2,k}$, which was normally fixed at 0. That parameter was freed for the 11-12 year age pool, with significant improvement to the three-factor model (see Table 31). Large partial derivatives were often obtained for Similarities' PO loading ($\lambda_{2,2,k}$) and Picture Completion's VC loading ($\lambda_{7,1,k}$) from tests of Kaufman's three-factor model for various age groups. These parameters were freed with significant improvement to the three-factor model for the total sample (see Table 14) and the 13-14 year age pool (see Table 36). The resultant loadings were very small, although significant. The correlation between Picture Completion and Similarities did not account for the Verbal factor loading of Picture Completion in the 9-10 year age pool. In that sample, Picture Completion was more highly correlated with all the Verbal Scale subtests, with the exception of Digit Span, than with any of the other Performance Scale subtests (see Table 10). The residual correlations between Similarities and Picture Completion contributed a substantial amount of noise to the clinical models for the two oldest age pools. These correlations are particularly interesting in view of the finding that Similarities tended to be extremely positively skewed at lower age levels, while P.C tended to be negatively skewed at various age levels.

The source(s) of the residual correlations involving Similarities and Picture Completion are not directly detectable from the configuration of the respective scaled scores or the analysis in the present study. Kaufman (1979a) has suggested that both of these subtests require the SI operation of Cognition (Guilford, 1967) and are measures, in part, of the skill of distinguishing essential from nonessential details. Similarities and Picture Completion differ in that verbal and spatial processing demands are attributed to them, respectively. Similarities differs from all other subtests in the demand for categorical, or "logical abstractive" (Kaufman, 1979a, p. 103) thinking which is attributed to that subtest. In sum, Picture Completion is thought to require the child to scan a pictoral stimuli and detect a detail which is conspicuous by its absence, i.e., whose absence makes the picture somehow incongruous. For example, the absence of a watchband on a drawing of a man's wrist is incongruous with the presence of a watchface on the wrist. In contrast, Similarities is thought to require the child to respond to the auditory representation of a pair of stimuli, such as an apple and a banana, and detect the most abstract feature which is common to both stimuli. For example, identification of the fact that both apples and bananas are fruits is worth two points, while the less abstract response that both grow on trees is awarded only a single point. The detail to be sought is one which defines a concept, such as fruits. At first glance, the above description of these subtests appears to indicate that the residual covariance of Similarities and Picture Completion is a function of the degree to which they require discrimination of essential and nonessential details, while their opposing distribution properties reflect deficits in verbal and/or concept attainment abilities, which would effect only Similarities scores. However, there are features of the administration and scoring of Similarities which could negatively bias the score of a child who knew the concepts represented in the Similarities items, but had a limited English vocabulary and limited test-wiseness.

As stated above, there are two types of responses to Similarities items which may be scored as correct. Responses which describe salient common features of the stimulus pair are awarded one point; those responses which identify the shared concept or category by name, two points. The greater abstraction attributed to the latter type of response is assumed to reflect more advanced cognitive development (Wechsler, 1974). An examination of the scoring guide for the subtest indicates that English vocabulary is a more salient feature in differentiating some one- and two-point responses. For example, "ways of communicating" is a two-point response to the stimulus pair "telephone-radio", while "transmit messages" is only a one-point response.

"Emotions" and "both feelings" are two-point responses to anger-joy, while "the way you feel" is a one-point response. "Utensils" is a two-point response to scissors-copper pan, while "household tools" is a one-point response. Children who have not been steadily exposed to English until they enter school might conceivably acquire an understanding of the concepts of utensils or emotions long before those words entered their vocabulary. MacArthur (Note 5) has described testing sessions with Inuit subjects who were unable to give the definition of "knife", although those subjects had demonstrated considerable skill with their own knives in various tasks outside the testing situation. Lacking knowledge of the English classificatory label, all such subjects can do is attempt to describe the concept with phrases such as "the way you feel", for which they receive a single point. As with Schubert and Cropley's (1972) Cree-speaking northern Saskatchewan sample, standard inferences about the level of mental processing involved in making a correct response may be biased by the subject's inability to express, at least in English, the rules and processes used to arrive at that response.

Responses on the Similarities subtest may also be affected by a feature of the organization of the items. The first four items are scored as 1 or 0, so that a response which lists details common to the stimulus pair, with no classification label, receives full points. If a child responds incorrectly to either of the first two items, the examiner provides an example of a correct response, according to a standard procedure described by Wechsler (1974). These example responses list common features or functions, rather than classifications. When higher scores for more abstract responses are introduced with the fifth item, the examiner is instructed by the test manual to provide the child with a mental set for such responses. If the child gives a single-point response to the fifth or sixth item, the examiner acknowledges the response as correct and then provides an example of a two-point response. The example is intended to serve as a signal to the child to respond with abstract categories. This signal appears to be quite subtle, particularly since the example of a two-point response is immediately preceded by verbal reinforcement for the child's single-point response. It may be too subtle for a child who lacks extensive experience with educational testing or who is struggling with English words in the test items.

The scoring procedures for Similarities may negatively affect the scores of Inuit children by invoking a theoretical assumption of tenuous validity. As described above, these procedures imply the assumption that categorical responses reflect greater capacity for abstraction. However, Kagan, Moss, and Sigel (1963) have demonstrated that the tendency to group figural stimuli according to a particular criterion, whether shared details within the stimuli or shared categories, is a cognitive style variable. People who are capable of grouping stimuli according to both modes will tend to prefer either the analytic (by detail) or categorical style. Kagan et al. suggested that the analytical style was associated with field independence, based on similar sets of relationships with ability and personality variables. This link would suggest that field independent children might approach Similarities items with a tendency to respond in an analytic style, which would result in fewer points. If the trends toward field independence in Inuit samples, which were observed by Berry (1966,1976) and MacArthur (1967, 1968, 1969, 1975a, 1975b) in the data they collected during the 1960's, could be replicated with current samples, then an additional source of bias against Inuit children may have been identified in the Similarities subtest.

The above discussion on possible biases in the Similarities subtest is speculative. Jensen (1980) has demonstrated the pitfalls involved in armchair attributions of test bias. However, given the organization of Similarities, failure to understand the subtest instructions or its implicit demands is a plausible explanation of the positive skewness of scores on that subtest for the three youngest age groups. This explanation is a viable alternative to attributions of deficits in concept formation, classification, or verbal ability per se among young Inuit children. The possibility of a predisposition to analytical responses, as a function of the same socialization practices thought to foster field independence, is a much less parsimonious explanation, but worth testing in view of the small, significant PO loading obtained for Similarities in the 13-14 year age pool results.

The above hypotheses pertaining to sources of possible bias in Similarities could be partially tested with a series of studies involving Similarities, Picture Completion, and Kagan et al's (1963) Conceptual Style Test (CST). Each item on this test consists of the presentation of a stimulus card containing line drawings of three figures. The child is asked to select the two figures which are alike. The sets of figures are organized to allow groupings based on at least two styles of response. Each item response is scored as analytic, categorical, or relational (where the figures selected have a functional relationship, e.g. an actor-to-object relationship). Cross-tabulations of CST response styles with scores on Similarities items would provide an index of the degree to which preferences on the analytic/categorical dimension generalize to Similarities items. Flexibility of response style on either test could be measured as a child's ability to switch styles following explicit instructions to that effect or verbal reinforcement of the nonpreferred style. It might be informative to devise new items for each of Similarities and the CST, such that the analytical, categorical, and functional relationships depicted in an item on the latter test would correspond to two-point, single-point, and failing responses on a parallel Similarities-type item. Flexibility of response style on each of the two tests could be compared under various experimental manipulations of administration conditions. Understanding of English nouns corresponding to two-point responses for the Similarities-type items should be insured. If pictorial, CST-type items were designed to depict the relationships represented in WISC-R Similarities items, children could be administered Similarities and its pictorial parallel in counterbalanced order. If children were able to respond in a categorical style on the CST-type test, but failed to make abstract two-point responses on the corresponding Similarities items, this result would be evidence that the low scores on Similarities reported in the literature, and the pattern of distributional properties and factor analytic results in the present samples, were affected by stimulus properties of the subtest.

The research proposals listed briefly above are offered as suggestions for examining the source of the residual correlation between Similarities and Picture Completion, which appeared to be related to deviations from Kaufman's model in the two older age pools. The results of such studies might further affect estimation of the construct validity and/or the clinical utility of the WISC-R for children in the Keewatin and Kitikmeot regions.

The positively-skewed distributions obtained for Similarities in the younger age pools probably reflects a level of English fluency which was too low for the test instructions to be understood. As mentioned in Chapter III.C, 20 children aged 7 and 8 years were excluded from the standardization analysis due to Mulcahy and Watters' (1982) determination that their English skills were not sufficient for valid assessment. Seven children of age 9 to 10 years were similarily excluded. Of those children remaining, 30 in the 7-8 year age pool obtained scores of 0 on Similarities. Of the children retained in the 9-10 year age pool, 25 obtained Similarities scores of 0. Many of the children from the youngest age pool appeared confused by the subtest questions and remained silent. Others replied that the stimulus pairs in question were not alike. Given the limited experience in English which is common for Inuit children in this age range, failure to understand the instructions appears to be the most parsimonious explanation for item failures on the early items. The recommendation to abstain from using the WISC-R with NWT children younger than 11 years of age is supported by the evidence for such difficulties.

Digit Span

At the conclusion of Chapter IV.C, several conclusions and hypotheses regarding the factor definition(s) of Digit Span were presented. The finding that the forward (DSF) and backward (DSB) sections of the subtest were significantly correlated and loading together in three-factor solutions for most age groups led to the conclusion that the tests do reflect a common ability construct. The relatively small values of those correlations, plus the finding that DSF occasionally loaded on a factor separate from DSB in the two-factor solutions or had no salient loadings, led to the conclusion that Digit Span was not an internally consistent subtest and that separate factors were contributing to the variance of DSF and DSB. The hypothesis that DSF measured verbal memory, perhaps aided by rehearsal, whereas DSB measures visual memory span, where the children read the digit string from a visual mental representation, was not supported by the factor analysis of DSF, DSB, and the remaining 11 WISC-R subtests. The hypothesized split of DSF and DSB to the Verbal and Performance factors, respectively, was only replicated in the results of the 8 year old age group. For the older groups, the two components of Digit Span (particularly DSB) loaded on a Performance factor or a Sequencing factor,depending on the number of factors rotated. These findings and conclusions led to a series of hypotheses about the nature of the factor(s) shared by DSF and DSB, and those factors which affect the two components differentially.

The issues of fluency in English and comprehension of the subtest instructions is relevant to the interpretation of Digit Span. The loading of Digit Span on Verbal in the ML analysis for only the 7-8 year age pool lends some support to this hypothesis. More concrete evidence was provided by the response of some children in this age pool to the initial DSF items, as observed by the present author during testing. Upon presentation of the first string of digits, several children responded by saying a number equal to the correct sum of the three digits. This response style was repeated on the second three-digit trial. As required by the administration rules for the subtest, DSF was then terminated and the child received a score of 0 on DSF. However, the administration of DSB includes a practice trial with a three-digit string. If the child answers incorrectly, the correct response is provided by the experimenter and a second practice trial is given. All of the children who had responded to DSF items with correct sums, responded in the same manner to the first practice trial of DSB but responded correctly to the second practice trial and subsequent items of the test. When errors were finally committed, they were of the digit reversal, substitution, or exclusion categories which are typically committed on this subtest. Following the completion of DSB, the present author asked these children to "do a few more forward". All of the children were then able to reproduce a forward span which was longer than their longest backward span.

The present author's interpretation of the above sequence is as follows. These children did not have sufficient fluency in English to understand the instructions for DSF. They had been in school long enough to make a reasonable guess as to what was demanded of them upon presentation of a string of numbers, i.e., to add them. (Teachers in NWT informed the present author that children in their schools were relatively skilled in arithmetic.) When provided with the correction of this false assumption during administration of DSB, the children realized the true demands of the task and responded accordingly. The word "forward" is not included in the standardized instructions for DSF. The word "backward" is emphasized in the instructions for DSB and repeated twice if the correction is given. When asked to continue the forward items, understanding of the words "forward" and "backward", plus the context of the testing in DSB, may have been sufficient information for the child to understand the new task demands. As noted in Chapter II.C, responses to certain directives, such as instructions to move forward and backward, become routinized for children in school and may lead adults to make the false assumption that more complex or ambiguous English phrases are understood.

The above sequence of Digit Span responses did not occur for all children in the 7-8 year age pool. This sequence occurred with more frequency in the less populated of the two settlements in which the present author administered the WISC-R. Watters (Note 6) has indicated that such occurrences are not uncommon in administration of the test battery to NWT children. Familiarity with English may have affected Digit Span scores in additional ways. Given the finding that speed of recognition of the stimulus items contributes to much of the variance on digit span tasks (Dempster, 1981), and demonstrations that the digit span length of bilingual subjects may vary with differences in the language of stimulus presentation (first or second language) (Ellis & Hennelly, 1980; Hoosain, 1982), fluency in English may have contributed to quicker recognition of Digit Span items and hence to increased memory capacity for some NWT subjects. It would follow from this hypothesis that the trends to load on the Performance factor in the two-factor solutions for older age groups (for Digit Span and for DSF and DSB), suggest that the digit names are sufficiently overlearned by most children that variance in fluency of English is not associated with variance in length of span. Comparison of correlations among speed of recognition of the digit names, English vocabulary, and Digit Span scores across age groups would provide a test of this hypothesis. The correlations of English vocabulary with both recognition speed and Digit Span scores would be expected to decrease with increasing age. This type of study would ideally be conducted longitudinally with several cohorts.

Although the correlations obtained between DSF and DSB tended to be relatively low, they tended to load together on the third factor of three-factor solutions. This finding indicates that while the subtest should not be interpreted as a homogenous measure of short term memory (or any other single ability), joint contributions to the third factor appear to be justified. At present, evidence from confirmatory factor analytis with MLFA is only available for inclusion of the scaled score for the total Digit Span subtest. Therefore, this score should be used for calculation of scores on the third factor, rather than separate weightings for DSF and DSB scores.

Given that the two-factor models were rejected for the two oldest age pools and use of the test is discouraged for the others, the Performance loadings of Digit Span in the two-factor solutions for ages 9 years to 14 years may not be worthy of further investigation. However, the relationship of both DSF and DSB to scores on visual sequential memory and scanning tasks would provide some additional information related to the theory that children respond to DSB by reading digits from a mental visual representation. The Performance loadings of both DSF and DSB indicate that older children in the sample may have been using a visual scanning strategy for both parts of the subtest. Since mental representations are not directly manipulable, correlational and experimental studies with visual test stimuli are probably the closest approximation available. Given the difficulties which young NWT children experienced with the WISC-R Digit Span, assessment of sequential memory should include visual tasks anyway.

Picture Arrangement and Mazes

The inconsistency of factor loadings for Picture Arrangement and Mazes across the age pools led to the conclusion that these subtests should be excluded from calculation of factor scores. If a clear developmental trend had emerged in the factor loadings, such as occurred with Digit Span, this trend may have been incorporated into interpretation of the factors which may underlie variance on subtest scores.

The Verbal and VC factor loadings for Picture Arrangement are not inconsistent with results obtained in U.S. samples, including the national standardization sample (Kaufman, 1975). Kaufman (1979a) lists Verbal Comprehension as a secondary factor underlying this subtest. Skills thought to contribute to performance on Picture Arrangement include planning ability, visual perception and organization, nonverbal reasoning, social judgement and the ability to distinguish essential and nonessential details (Kaufman, 1979a). Given this large array of hypothesized component abilities, and the inconsistency of factor definitions for the subtest in the present sample, exclusion of the subtest from the calculation in the PO factor seems justified.

The Sequencing loading which was obtained for Mazes with the 11-12 year sample lends some face validity to the theory that the factor measures ability in sequential processing. Naglieri, Kamphaus, and Kaufman (1983) had hypothesized that Mazes would load on this factor when they attempted to validate their simultaneous-successive processing interpretation of the WISC-R. However, construct interpretation of the third factor requires the sort of multivariate experimental research advocated by Messick (1972), in which instructions, task demands and materials, and skills on component processes are varied to determine the effect upon factor patterns and factor scores.

D. Limitations of the Study

The present study was designed to test the construct validity of the WISC-R for children in the Keewatin and Kitikmeot regions by attempting to replicate the factor models implied by three sets of guidelines for interpreting clusters of WISC-R subtests. The results were not intended to provide valid psychological interpretations of the factors produced. In Embretson's (1983) terminology, this study investigated the nomothetic span of the test, and not the construct representation of the subtests or factors. Nontheless, there are aspects of the design and execution of the analysis which limit the reliability and generalizability of the results. These are discussed below.

The final solutions which were presented as the best factor models for the various age pools must be considered as tentative until replicated in an independent sample from the same populations. This restriction is based on Cliff's (1983) argument that models adjusted according to results of previous analyses of a given set of data lose their status as hypotheses when tested against the same data. Given the investment of time and resources necessary for the collection of similar data, this replication will probably not be attempted.

The size of the age group and age pool samples is smaller than the number desired for such a study. The ratio of subjects to variables ranges from 8.0 to 8.58 for the three oldest age pools. This is not far below Nunnally's (1978) recommendation of 10 subjects per variable. All age pool sample sizes exceeded Lawley and Maxwell's (1971) rule of thumb that the difference between the sample size and the number of variables exceed 50. The subjects-to-variable ratio of 5.92, which was obtained for the 7-8 year age pool, should lead to further caution in interpretation of the results for that sample. However, there was still sufficient power to reject the clinical models for the younger age groups. Since these decisions were based in part on the intrusion of salient loadings which had been originally set to 0, and not simply on the nonsignificance of unconstrained loadings, the clinical models appear to have been rejected in spite of the small sample sizes, rather than as a result of them.

The decision to lend more credence to the separate age pool results than the total sample results involved human judgement in the setting of criteria for that decision. The statement of Hypothesis 14 in Chapter II.E declared that validation of the total sample results to an age pool required that the cross-validation χ^2 be nonsignificant and that the free parameters in the model be significant. The age pool results were given more weight due to the failure of the total sample's three-factor pattern to meet the latter criteria. Had the decision been based on arguments of statistical robustness alone, the analysis of the total sample correlation matrix would have been meaningless (and therefore dishonest), since the difficulties posed by differences in sample sizes were known before analysis began. The present author would then have been guilty of the same subjective procedures for which he has criticized other authors (e.g., Bejar & Doyle, 1981; Guilford and Hoepfner, 1971; McGaw & Jöreskog, 1971). Although the decision was based on an a priori set of criteria, the validity of the criteria is open to question. Their definition was partially based on the present author's previously-stated assumption that the onus for validating the test, for each intended purpose and with each prospective population, lies with the users or providers of the test. The criteria were therefore defined so as to increase the likelihood of detecting age differences which might be present in the populations, i.e., to reduce Type II error. It could be argued that cross-validating the results of a large sample against the data for a subsample would almost always result in nonsignificant free parameters, given the decreased robustness for the subsample. As described in the summary of cross-validation results in Chapter IV.B and in the opening paragraph of Chapter V.B, there is a good deal of evidence to refute the claim that the age differences evident in the results for the present study are a spurious effect of sample size. However, more efficient and objective ways of reducing Type II error may have been, or may become, available.

The validity of conclusions regarding age trends is also limited by the cross-sectional nature of the study. Strong conclusions about age trends would require a combination of longitudinal and cross-sectional comparisons. Political and industrial developments in NWT in the near future may radically alter the delivery of education in the region. Age trends identified in the present study may reflect differential effects of such change on various cohorts. One example of such an effect is provided in the hypothesis that the introduction of television may have improved the school readiness skills of children in the youngest age pool, leading to factor analytic results which were less discrepant from the clinical models than the results of children aged 9 years to 10 years. Changes of this sort could also lead to the rapid obsolescence of Mulcahy and Watter's (1982) tables for the derivation of scaled scores.

The reliability and generalizability of the WISC-R scores may have been threatened by a high frequency of episodes of otitis media, which has been shown to affect the cognitive test scores of samples of Native children, particularly in the arctic regions. The proportion of subjects with histories of otitis media is not known. Children whose scores were excluded from analysis on the basis of numerous raw scores of 0 or lack of response may have been suffering an episode of ear infection or may have already sustained a permanent severe hearing deficit. The remaining children may also have been experiencing hearing difficulties during the testing session. In fact, such hearing deficits comprise an explanation for disturbances in factor models to rival that of familiarity with English, although it would be difficult to explain the acceptance of Kaufman's model for older children on the basis of the former explanation. Even if medical histories of the children were incorporated, this information would not have identified those children who were suffering an episode of infection during testing.

The children tested by the present author appeared to be highly motivated to complete the tests. Conversation with other testers indicated that this observation was made in all the communities. Preadolescents appeared to be more eager than adolescents, although the latter were generally very cooperative. Many young children who were not selected for the sample expressed disappointment to the testers and/or their teachers. The subtests which the children appeared to enjoy the most were Block Design and Object Assembly, followed by the remaining Performance Scale subtests. This pattern of preferences is similar to that observed among urban southern children by the present author. There was no clear indication that test results were affected by either low motivation or the presence of an unknown tester to a degree not found with children in southern Canada.

E. Summary Statement

The four conclusions stated at the beginning of this chapter may be summarized by stating that whereas construct validity for factor models of the WISC-R was not demonstrated for children younger than age 11 years, a three-factor model similar to Kaufman's was validated for children of 11 years to 14 years. The constructs represented by the factors are not known, i.e., although the subtests can be clustered into the hypothesized factors in a way which accounts for much of the subtest variances and covariances, the abilities measured by those factors cannot be determined from the present study. Research designed to address the question of construct representation has been proposed throughout this chapter. If the WISC-R is used for clinical purposes in NWT, such further research is essential. As discussed in chapter II.B, the identity of the constructs of Kaufman's factors is not clear for application to U.S. children. Given the number of factors which have been demonstrated to affect the scores of rural Native children whose first language is not English, the labels for

Kaufman's factors are even more tenuous for application to the present NWT sample.

The providers of psychological services in NWT should seriously question the value of sorting children on the basis of scores on these factors. Early diagnosis of learning difficulties would not be valid, given the failure to replicate clinical models for young children. This application of the WISC-R has been generally discredited anyway, as demonstrated in Chapter II.B. Factor scores may provide information for instructional design for older children, but this remains to be demonstrated. Given the findings of the present study and the limited utility of the ability training, or Construct, model of assessment for planning effective educational programs (Ysseldyke and Mirken, 1982), special educators in NWT are advised to develop and emphasize more funtional approaches to assessment.

Watters (in press) indicated a need for research on the learning styles and strengths of Native children. Further cognitive assessment research in NWT should address learning processes directly, i.e, examine changes in performance as a function of such variables as mode of instruction, test instructions, characteristics of the task materials, practice, or examiner prompts. The effect of manipulation of such variables on factor patterns and factor scores derived from normative test batteries would also be of interest Whatever instruments are eventually employed by special educators and psychologists in NWT, the educational needs of the children will not be served by repeating the errors of overinterpretation of test scores which have characterized the assessment of Native children to date.

Reference Notes

1. Kaufman, A. S. *Intelligent testing*. Paper presented at Clinical Symposium, University of British Columbia Education Clinic, Vancouver, B. C., 1982.
2. Mueller, H. H. *Bannatyne recategorized WISC-R patterning of normal and exceptional children: A meta-exploration*. Unpublished manuscript, University of Alberta, 1983.
3. Mueller, H. H., Mancini, G., & Short, R. H. *An evaluation of the diagnostic efficiency of the Wechsler Intelligence Scale for Children-Revised*. Manuscript submitted for publication, 1983.
4. Jöreskog, K. G. Personal communication, April 5, 1983.
5. MacArthur, R. S. Personal communication, September, 1979.
6. Watters, B. Personal communication, March, 1983.

References

Ackerman, P. T., Dykman, R. A., & Peters, J. E. Hierarchical factor patterns on the WISC as related to areas of learning deficit. *Perceptual and Motor Skills*, 1976, *42*, 583-615.

Advisory Committee on Northern Development. *1980-1981 Government activities in the north*. Ottawa: Department of Indian Affairs and Northern Development, 1981.

Alwin, D. F., & Jackson, D. J. Applications of simultaneous factor analysis to issues of factorial invariance. In D. J. Jackson & E. F. Borgatta (Eds.), *Factor analysis and measurement in sociological research*. London: Sage Publications, 1981.

American Psychological Association. *Standards for educational and psychological tests*. Washington, D. C.: American Psychological Association, 1974.

Applebaum, A. S., & Tuna, J. M. The relationship of the WISC-R to academic achievement in a clinical population. *Journal of Clinical Psychology*, 1982, *38*, 401-405.

Armstrong, J. S., & Soelberg, P. On the interpretation of factor analysis. *Psychological Bulletin*, 1968, *70*, 361-364.

Arter, J. A., & Jenkins, J. R. Differential diagnosis - prescriptive teaching: A critical appraisal. *Review of Educational Research*, 1979, *49*, 517-555.

Bakan, D. The test of significance in psychological research. *Psychological Bulletin*, 1966, *66*, 423-437.

Banas, N. & Wills, I. H. Prescriptions from WISC-R patterns. *Academic Therapy*, 1978, *13*, 491-495.

Bannatyne, A. Diagnosing learning disabilities and writing remedial prescriptions. *Journal of Learning Disabilities*, 1968, *1*, 242-248.

Bannatyne, A. Diagnosis: A note on recategorization of the WISC-R scaled scores. *Journal of Learning Disabilities*, 1974, *7*, 272-273.

Baumeister, A. A., & Bartlett, C. J. A comparison of the factor structure of normals and retardates on the WISC. *American Journal of Mental Deficiency*, 1962, *66*, 641-646. (a)

Baumeister, A. A., & Bartlett, C. J. Further factorial investigations of WISC performance of mental defectives. *American Journal of Mental Deficiency*, 1966, *67*, 257-261. (b)

Bechtoldt, H. P. Construct validity: A critique. *American Psychologist*, 1959, *14*, 619-629.

Bejar, I. I., & Doyle, K.O. Factorial invariance in student ratings of instruction. *Applied Psychological Measurement*, 1981, *5*, 307-312.

Bentler, P. M., & Bonett, D. G. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 1980, *88*, 588-606.

Berk, R. A. The value of WISC-R profile analysis for the differential diagnosis of learning disabled children. *Journal of Clinical Psychology*, 1983, *39*, 133-136.

Berry, J. W. Temne and Eskimo perceptual skills. *International Journal of Psychology*, 1966, *1*, 207-229.

Berry, J. W. On cross-cultural comparability. *International Journal of Psychology*, 1969, *4*, 119-128.

Berry, J. W. Psychological research in the north. *Anthropologica*, 1971, *13*, 143-157.

Berry, J. W. Ecological and cultural factors in spatial perceptual development. In J. W. Berry & P. R. Dasen (Eds.), *Culture and cognition: Readings in cross-cultural psychology*. London: Methuen, 1974.

Berry, J. W. *Human ecology and cognitive style: Comparative studies in cultural and psychological adaptation*. New York: Wiley, 1976.

Bland, L. L. *Perception and visual memory of school-age Eskimos and Athabascan Indians in Alaskan villages*. Human Environmental Resources Systems, 1970. (ERIC Document Reproduction Service No. ED 046 549)

Bland, L. L. *Visual perception and recall of school-age Navajo, Hopi, Jicarilla Apache, and Caucasian children of the southwest including results from a pilot study among Eskimos and Athabascan school-age children of North Alaska*. Kennewick, Wash.: Human Environmental Resources Services, 1975. (ERIC Document Reproduction Service No. ED 160 256)

Block, N.J., & Dworkin, G. IQ, heritability, and inequality. In N. J. Block & G. Dworkin (Eds.), *The IQ controversy*.

New York: Pantheon, 1976.

Bloom, A. S., & Raskin, L. M. WISC-R verbal-performance IQ discrepancies: A comparison of learning disabled children to the normative sample. *Journal of Clinical Psychology*, 1980, *36*, 322-323.

Bowd, A. D. Some determinants of school achievement in several Indian groups. *Alberta Journal of Educational Research*, 1972, *18*, 69-76.

Bowd, A. D. A cross-cultural study of the factorial composition of mechanical aptitude. *Canadian Journal of Behavioral Science*, 1973, *5*, 13-23.

Bowd, A. D. Linguistic background and nonverbal intelligence: A cross-cultural comparison. *Journal of Educational Research*, 1974, *68*, 26-27. (a)

Bowd, A. D. Practical abilities of Indians and Eskimos. *Canadian Psychologist*, 1974, *15*, 281-290. (b)

Bowd, A. D. Ten years after the Hawthorn Report: Changing psychological implications for the education of Canadian Native peoples. *Canadian Psychological Review*, 1977, *18*, 332-345.

Bowd, A. D., McDougall, D., & Yewchuk, C. *Ed. Psych.: A Canadian perspective*. Toronto: Gage, 1982.

Brandes, P. J., & Ehinger, D. M. The effects of early middle ear pathology on auditory perception and academic achievement. *Journal of Speech and Hearing Disorders*, 1981, *46*, 301-307.

Brady, P.M., Manni, J. L., & Winikur. A three-tiered model for the assessment of culturally and linguistically different children. *Psychology in the Schools*, 1983, *20*, 52-58

Brody, E. B., & Brody, N. *Intelligence: Nature, determinants, and consequences*. New York: Academic Press, 1976.

Brody, H. Eskimo: A language with a future? *Polar Research*, 1977, *18*, 587-592.

Browne, M. W. A comparison of factor analytic techniques. *Psychometrika*, 1968, *33*, 267-334.

Buss, A. R., & Royce, J. R. Detecting cross-cultural commonalities and differences: Intergroup factor analysis. *Psychological Bulletin*, 1975, *82*, 128-136.

Carney, R. E., & Trowbridge, N. Intelligence test performance of Indian children as a function of type of test and age. *Perceptual and Motor Skills*, 1962, *14*, 511-514.

Carpenter, E. Space concepts of the Eskimos. *Explorations*, 1955, *5*, 131-145.

Carroll, J. B. Remarks on Sternberg's "Factor theories of intelligence are all right almost". *Educational Researcher*, 1980, *9*(8), 14-18.

Carroll, J. B. Ability and task difficulty in cognitive psychology. *Educational Researcher*, 1981, *10*(1), 11-21. (a)

Carroll, J. B. Review of *Bias in Mental Testing* by A. R. Jensen. *Psychometrika*, 1981, *46*, 227-233. (b)

Cattell, R. B. *The scientific use of factor analysis in behavioral and life sciences.* New York: Plenum Press, 1978.

Cattell, R. B. & Baggaley, A. R. The salient variable similarity index for factor matching. *British Journal of Statistical Psychology*, 1960, *13*, 33-46.

Cattell, R. B., Balcar, K. R., Horn, J. L., & Nesselroade, J. R. Factor matching procedures: An improvement of the s index; with tables. *Educational and Psychological Measurement*, 1969, *29*, 781-792.

Clarizio, H., & Bernard,R. Recategorized WISC-R scores of learning disabled children and differential diagnosis. *Psychology in the Schools*, 1981, *18*, 5-12.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. Educational uses of tests with disadvantaged students. *American Psychologist*, 1975, *30*, 15-41.

Cliff, N. Orthogonal rotation to congruence. *Psychometrika*, 1966, *31*, 33-42.

Cliff, N. Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, 1983, *18*, 115-126.

Coates, S. Field independence and intellectual functioning in preschool children. *Perceptual and Motor Skills*, 1975, *41*, 251-254.

Cohen, J. The factorial structure of the WISC at ages 7-6, 10-6, and 13-6. *Journal of Consulting Psychology*, 1959, *23*, 285-299.

Cole, M., & Bruner, J. Cultural differences and inferences about psychological processes. *American Psychologist*, 1971, *26*, 867-876.

Comrey, A. L. Common methodological problems in factor analytic studies. *Journal of Consulting and Clinical Psychology*, 1978, *46*, 648-659.

Connelly, J. B. Recategorized WISC-R score patterns of older and younger referred Tlingit Indian children. *Psychology in the Schools*, 1983, *20*, 271-275.

Cooper, J. C. B. Factor analysis: An overview. *American Statistician*, 1983, *37*, 141-147.

Costa, L. D. The relation of visuospatial dysfunction to Digit Span performance in patients with cerebral lesions. *Cortex*, 1975, *11*, 31-36.

Covin, T. M., & Lubimiv, A. J. Concurrent validity of the WRAT. *Perceptual and Motor Skills*, 1976, *43*, 573-574.

Cress, J. N. Cognitive and personality testing use and abuse. *Journal of American Indian Education*, 1974, *13*(3), 16-19.

Cronbach, L. J. Five decades of controversy over mental testing. *American Psychologist*, 1975, *30*, 1-14.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, *52*, 281-302.

Cummins, J. Educational implications of mother tongue maintenance in minority language groups. *The Canadian Modern Language Review*, 1978, *39*, 395-416.

Cummins, J. Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, 1979, *49*, 222-251.

Cummins, J. Psychological tests and ESL students. *TEAL 81 - TESL Canada conference proceedings*, 1982, *1*, 63-66.

Cummins, J. P., & Das, J. P. Cognitive processing, academic achievement, and WISC-R performance in EMR children. *Journal of Consulting and Clinical Psychology*, 1980, *48*, 777-779.

Cunningham, W. R. Principles for identifying structural differences: Some methodological issues related to comparative factor analysis. *Journal of Gerontology*, 1978, *33*, 82-86.

Cunningham, W. R. Ability factor structure differences in adulthood and old age. *Multivariate Behavioral Research*, 1981, *16*, 3-22.

Darcy, N. A review of the literature on the effects of bilingualism upon the measurement of intelligence. *Journal of Genetic Psychology*, 1953, *82*, 21-57.

Darcy, N. Bilingualism and the measurement of intelligence: Review of a decade of research. *Journal of Genetic Psychology*, 1963, *103*, 259-282.

Darlington, R. B. Another look at cultural fairness. *Journal of Educational Measurement*, 1971, *8*, 71-82.

Das, J. P., Kirby, J. R., & Jarman, R. F. *Simultaneous and successive cognitive processes*. New York: Academic Press, 1979.

Das, J. P., & Krywaniuk, L. W. *Memory and reasoning in Native children: An effort at improvement through the teaching of cognitive strategies*. Edmonton, Ab.: University of Alberta, 1972.

Das, J. P., Manos, J., & Kanungo, R. N. Performance of Canadian native, black and white children on some cognitive and personality tests. *Alberta Journal of Educational Research*, 1975, *21*, 183-195.

Dean, R. S. Reliability of the WISC-R with Mexican-American children. *Journal of School Psychology*, 1977, *15*, 267-268.

Dean, R. S. Factor structure of the WISC-R with Anglos and Mexican-Americans. *Journal of School Psychology*, 1980, *18*, 234-239.

Dean, R. S. Intelligence as a predictor of nonverbal learning with learning-disabled children. *Journal of Clinical Psychology*, 1983, *39*, 437-441.

Dempster, F. N. Memory span: Sources of individual and developmental differences. *Psychological Bulletin*, 1981, *89*, 63-100.

Dempster, F. N., & Zinkgraph, S. A. Individual differences in Digit Span and chunking. *Intelligence*, 1982, *6*, 201-213.

Denny, J. P. Notes on thinking processes facilitated by the Eskimo and English languages. *Musk-ox*, 1973, *12*, 81.

Devine, M., & Wood, D. *NWT data book 1981*. Yellowknife,NWT:

Outcrop, 1981.

DIAND. *Indian conditions: A survey*. Ottawa: Author, 1980.

Dillon, R. F., & Stevenson-Hicks, R. Competence vs. performance and recent approaches to cognitive assessment. *Psychology in the Schools*, 1983, *30*, 142-145.

Downing, J., Ollila, L., & Oliver, P. Cultural differences in children's concepts of reading and writing. *British Journal of Educational Psychology*, 1975, *45*, 312-316.

Drenth, P. J. D. Implications of testing for individual and society. In L. J. Cronbach & P. J. D. Drenth (Eds.), *Mental tests and cultural adaptation*. The Hague: Mouton, 1972.

Driel, O. P. van. On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika*, 1978, *43*, 225-243.

Elleys, W. B., & MacArthur, R. S. The Standard Progressive Matrices as a culture reduced measure of general intellectual ability. *Alberta Journal of Educational Research*, 1962, *8*, 54-65.

Elliott, S. N., & Bretzing, B. H. Using and updating local norms. *Psychology in the Schools*, 1980, *17*, 196-201.

Ellis, N. C., & Hennelly, R. A. A bilingual word-length effect: Implications for intelligence testing and the relative ease of mental calculation in Welsh and English. *British Journal of Psychology*, 1980, *71*, 43-51.

Embretson, S. Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 1983, *93*, 179-197.

Evans, J. G. & Hamm, H. A. An argument for administering WISC-R Digit Span. *Perceptual and Motor Skills*, 1979, *49*, 573-574.

Feldman, C. F., & Bock, R. D. Cognitive studies among residents of Wainwright Village, Alaska. *Arctic Anthropology*, 1970, *7*(1), 101-108.

Finch, A. J. Jr. Kendall, P. C., Spirito, A., Entin, A., Montgomery, L. E., & Schwartz, D. J. Short form and factor analytic studies of the WISC-R with behavior problem children. *Journal of Abnormal Child Psychology*, 1979, *7*, 337-344.

Flaugher, R. L. The many definitions of test bias. *American*

Psychologist, 1978, *33*, 671-679.

Fleishman, E. A. Individual differences and motor learning. In R. M. Gagne (Ed.), *Learning and individual differences*. Columbus, OH.: Merrill, 1967.

Fleishman, E. A., & Hempel, W. E. Changes in factor structure of a complex psychomotor task as a function of practice. *Psychometrika*, 1954, *19*, 239-252.

Gaarder, A. B. *Bilingual schooling and the survival of Spanish in the United States*. Rowley, Mass.: Newbury House, 1977.

Gaddes, W., McKenzie, A., & Barnsley, R. Psychometric intelligence and spatial imagery in two northwest Indian and white groups of children. *Journal of Social Psychology*, 1968, *75*, 35-42.

Gagne, R. C. Spatial concepts in the Eskimo language. In V. F. Valentine & F. G. Vallee (Eds.), *Eskimo of the Canadian arctic*. Toronto: McClelland & Stewart, 1968.

Gall, P. *Early school leaving and school progress of native students in the North-West Territories: An exploratory study*. Unpublished thesis, University of Alberta, 1980.

Gardner, R. A. Digits Forward and Digits Backward as two separate tests: Normative data on 1567 school children. *Journal of Clinical Child Psychology*, 1981, *10*, 131-135.

Garrett, H. E. A. A developmental theory of intelligence. *American Psychologist*, 1946, *1*, 372-378.

Geweke, J. F., & Singleton, K. J. Interpreting the likelihood ratio statistic when sample size is small. *Journal of the American Statistical Association*, 1980, *75*, 133-137.

Glass, G., & Stanley, J. C. *Statistical methods in education and psychology*. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Goldman, R. D., & Slaughter, R. E. Why college grade point average is difficult to predict. *Journal of Educational Psychology*, 1976, *68*, 9-14.

Goodenough, D. R., & Karp, S. A. Field dependence and intellectual functioning. *Journal of Abnormal and Social Psychology*, 1961, *63*, 241-246.

Government of the Northwest Territories, Legislative Assembly. *Learning: Tradition and change in the Northwest Territories*. Yellowknife, NWT: Author, 1982.

Griffin, P. T. & Hefferman, A. Digit Span Forward and Backward: Separate and unequal components of the WAIS Digit Span. *Perceptual and Motor Skills*, 1983, *56*, 335-338.

Groff, M. G., & Hubble, L. WISC-R factor structures of younger and older youth with low IQs. *Journal of Consulting and Clinical Psychology*, 1982, *50*, 148-149.

Groff, M. G., & Linden, K. W. The WISC-R factor score profiles of cultural-familial mentally retarded and nonretarded youth. *American Journal of Mental Deficiency*, 1982, *87*, 147-152.

Grossman, F. M., & Johnson, K. M. WISC-R factor scores as predictors of WRAT performance: A multivariate analysis. *Psychology in the Schools*, 1982, *19*, 465-468.

Guilford, J. P. When not to factor analyze. *Psychological Bulletin*, 1952, *49*, 26-37.

Guilford, J. P. *The nature of human intelligence*. New York: McGraw Hill, 1967.

Guilford, J. P. Rotation problems in factor analysis. *Psychological Bulletin*, 1974, *81*, 498-501.

Guilford, J. P. The invariance problem in factor analysis. *Educational and Psychological Measurement*, 1977, *37*, 11-19.

Guilford, J. P. Cognitive styles: What are they? *Educational and Psychological Measurement*, 1980, *40*, 715-735.

Guilford, J. P., & Hoepfner, R. *The analysis of intelligence*. New York: McGraw-Hill, 1971.

Gutkin, T. B. Some useful statistics for the interpretation of the WISC-R. *Journal of Consulting and Clinical Psychology*, 1978, *46*, 1561-1563.

Gutkin, T. B. Bannatyne patterns of Caucasian and Mexican-American learning disabled children. *Psychology in the Schools*, 1979, *16*, 178-183. (a)

Gutkin, T. B. WISC-R scatter indices: Useful information for differential diagnosis? *Journal of School Psychology*, 1979, *17*, 368-371. (b)

Gutkin, T. B. The WISC-R Verbal Comprehension, Perceptual Organization, and Freedom from distractibility Deviation Quotients: Data for practitioners. *Psychology in the Schools*, 1979, *16*, 359-360. (c)

Gutkin, T. B. WISC-R Deviation Quotients vs. traditional IQs; an examination of the standardization sample and some implications for score interpretation. *Journal of Clinical Psychology*, 1982, *33*, 179-182.

Gutkin, T. B., & Reynolds, C. R. Factorial similarity of the WISC-R for Anglos and Chicanos referred for psychological services. *Journal of School Psychology*, 1980, *18*, 34-39.

Gutkin, T. B., & Reynolds, C. R. Factorial similarity of the WISC-R for white and black children from the standardization sample. *Journal of Educational Psychology*, 1981, *73*, 227-231.

Hale, R. The utility of WISC-R subtest scores in discriminating among adequate and underachieving children. *Multivariate Behavioral Research*, 1979, *14*, 245-253.

Hannah, G. S., House, B., & Salisbury, L. H. WAIS performance of Alaska native university freshmen. *Journal of Genetic Psychology*, 1968, *112*, 57-61.

Harman, H. *Modern factor analysis* (2nd. ed.). Chicago: University of Chicago Press, 1967.

Harris, D. B. *Children's drawings as measures of intellectual maturity*. New York: Harcourt, Brace, & World, 1963.

Hartlage, L. C., & Boone, K. E. Achievement test correlates of Wechsler Intelligence Scale for Children and Wechsler Intelligence Scale for Children - Revised. *Perceptual and Motor Skills*, 1977, *45*, 1283-1286.

Hartlage, L. C., & Steele, C. T. WISC and WISC-R correlates of academic achievement. *Psychology in the Schools*, 1977, *14*, 12-15.

Havighurst, R., Gunther, M., & Pratt, I. Environment and the Draw-A-Man test: the performance of Indian children.*Journal of Abnormal and Social Psychology*, 1946, *41*, 50-63.

Hays, W. L. *Statistics for psychologists*. New York: Holt, Rinehart, and Winston, 1964.

Hennessy, J. J., & Merrifield, P. R. A comparison of the factor structures of mental abilities in four ethnic groups. *Journal of Educational Psychology*, 1976, *68*, 754-759.

Hirshoren, A., & Kavale, K. Profile analysis of the WISC-R: A continuing malpractice. *The Exceptional Child*, 1976, *23*, 83-87.

Hodges, K. Factor structure of the WISC-R for a psychiatric sample. *Journal of Consulting and Clinical Psychology*, 1982, *50*, 141-142.

Hodges, W. E., & Spielberger, C. D. Digit Span:An indicant of trait or state anxiety. *Journal of Consulting and Clinical Psychology*, 1969, *33*, 430-434.

Hoosain, R. Correlation between pronunciation speed and Digit Span size. *Perceptual and Motor Skills*, 1982, *55*, 1128.

Horn, J. L. On subjectivity in factor analysis. *Educational and Psychological Measurement*, 1967, *27*, 811-820.

Horn, J. L. Organization of abilities and the development of intelligence. *Psychological Review*, 1968, *75*, 242-259.

Horn, J. L. On the internal consistency reliability of factors. *Multivariate Behavioral Research*, 1969, *4*, 115-125.

Horn, J. L. Trends in the measurement of intelligence. *Intelligence*, 1979, *3*, 229-240.

Horn, J. L., & Knapp, J. R. On the subjectivity of the empirical base of Guilford's structure-of-intellect model. *Psychological Bulletin*, 1973, *80*, 33-43.

Horn, J. L., & Knapp, J. R. Thirty wrongs do not make a right. *Psychological Bulletin*, 1974, *81*, 502-504.

Hubble, L. M., & Groff, M. Factor analysis of the WISC-R scores of male delinquents referred for evaluation. *Journal of Consulting and Clinical Psychology*, 1981, *49*, 738-739.

Hull, C. H., & Nie, N. H. *SPSS update 7-9: New procedures and facilities for releases 7-9*. New York: McGraw-Hill, 1981.

Humphreys, L. G. Statistical definitions of test validity for minority groups. *Journal of Applied Psychology*, 1973, *58*, 1-4.

Humphreys, L. G., Ilgen, D., McGrath, D., & Montanelli, R. Capitalization on chance in rotation of factors. *Educational and Psychological Measurement*, 1969, *29*, 259-271.

Humphreys, L. G., & Parsons, C. K. Partialling out intelligence: A methodological and substantive contribution. *Journal of Educational Psychology*, 1977, *69*, 212-216.

Humphreys, L. G., & Taber, T. Ability factors as a function of advantaged and disadvantaged groups. *Journal of Educational Measurement*, 1973, *10*, 107-115.

Hynd, G. W., & Garcia, W. I. Intellectual assessment of the Native American student. *School Psychology Digest*, 1979, *8*, 446-454.

Hynd, G. W., Quackenbush, R., Kramer, R., Conner, R., & Weed, W. Clinical utility of the WISC-R and the French Pictorial Test of Intelligence with Native American primary grade children. *Perceptual and Motor Skills*, 1979, *49*, 480-482.

James, L. R. Criterion models and construct validity for criteria. *Psychological Bulletin*, 1973, *80*, 75-83.

Jennrich, R. I. An asymptotic χ^2 test for the equality of two correlation matrices. *Journal of the American Statistical Association*, 1970, *65*(330), 904-912.

Jensen, A. R. Do schools cheat minority children? *Educational Research*, 1971, *14*, 3-28.

Jensen, A. R. Test bias and construct validity. *Phi Delta Kappan*, 1976, *58*, 340-346.

Jensen, A. R. *g*: Outmoded theory or unconquered frontier? *Creative Science and Technology*, 1979, *2*, 16-29.

Jensen, A. R. *Bias in mental testing*. New York: Free Press, 1980.

Jensen, A. R., & Figueroa, R. A. Forward and backward Digit Span interaction with race and IQ: Predictions from Jensen's theory. *Journal of Educational Psychology*, 1975, *67*, 882-893.

Jirsa, J. E. The SOMPA: A brief examination of technical considerations, philosophical rationale, and implications for practice. *Journal of School Psychology*, 1983, *21*, 13-21.

Johnson, E. G., & Lyle, J. G. Analysis of WISC Coding: 1. Figural reversibility. *Perceptual and Motor Skills*, 1972, *34*, 195-198. (a)

Johnson, E. G., & Lyle, J. G. Analysis of WISC Coding: 2. Memory and verbal mediation. *Perceptual and Motor*

Skills, 1972, *34*, 659-652. (b)

Johnson, E. G., & Lyle, J. G. Analysis of WISC Coding: 4. Paired-associate learning and performance strategies. *Perceptual and Motor Skills*, 1973, *37*, 695-698.

Jöreskog, K. G. On the statistical treatment of residuals in factor analysis. *Psychometrika*, 1962, *27*, 335-354.

Jöreskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 1969, *34*, 183-202.

Jöreskog, K. G. Simultaneous factor analysis in several populations. *Psychometrika*, 1971, *36*, 409-426.

Jöreskog, K. G. Structural analysis of covariance and correlation matrices. *Psychometrika*, 1978, *43*, 443-477.

Jöreskog, K. G. Analyzing psychological data by structural analysis of covariance matrices. In K. G. Jöreskog & D. Sörbom (Eds.). *Advances in factor analysis and structural equation models*. Cambridge, Mass.: Abt Books, 1979 (a)

Jöreskog, K. G. Author's addendum to "A general approach to confirmatory maximum likelihood factor analysis". In K. G. Jöreskog & D. Sörbom (Eds.). *Advances in factor analysis and structural equation models*. Cambridge, Mass.: Abt Books, 1979. (b)

Jöreskog, K. G. Basic ideas of factor and component analysis. In K. G. Jöreskog & D. Sörbom (Eds.), *Advances in factor analysis and structural equation models*. Cambridge, Mass.: Abt Books, 1979. (c)

Jöreskog, K. G. Basic issues in the application of LISREL. *Data*, 1981, *1*, 1-6.

Jöreskog, K. G., & Lawley, D. N. New methods in maximum likelihood factor analysis. *British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 85-96.

Jöreskog, K. G., & Sörbom, D. *LISREL IV: Analysis of linear structural relationships by the method of maximum likelihood*. Chicago: International Educational Services, 1978.

Jöreskog, K. G., & Sörbom, D. *LISREL V: Analysis of linear structural relationships by maximum likelihood and least squares methods*. Chicago: National Educational Resources, 1981.

Kagan, J., Moss, H., & Sigel, I. Psychological significance of styles of conceptualization. In J. C. Wright & J. Kagan (Eds.), *Basic cognitive processes in children.* Monographs of the Society for Research in Child Development, 1963, *28* (2, Serial No. 86).

Kaiser, H. F. A second generation Little Jiffy. *Psychometrika*, 1970, 235, 401-415.

Kaiser, H. F., Hunka, S., & Bianchini, J. C. Relating factors between studies based upon different individuals. *Multivariate Behavioral Research*, 1971, *6*, 409-422.

Kaplan, G. J., Fleshman, J. K., Bender, T.R., Baum, C., & Clark, P.S. Long-term effects of otitis media: a ten-year cohort study of Alaskan Eskimo children. *Pediatrics*, 1973, *52*, 577-585.

Karnes, F. A., & Brown, K. E. Factor analysis of the WISC-R for the gifted. *Journal of Educational Psychology*, 1980, *72*, 197-199.

Karp, S. Field dependence and overcoming embeddedness. *Journal of Consulting Psychology*, 1963, *27*, 294-302.

Katzenmeyer, W. G., & Stenner, A. J. Estimation of the invariance of factor structures across sex and race with implications for hypothesis testing. *Educational and Psychological Measurement*, 1977, 37,111-119.

Kaufman, A. S. Factor analysis of the WISC-R at 11 age levels between 6½ and 16½ years. *Journal of Consulting and Clinical Psychology*, 1975, *43*, 135-147.

Kaufman, A. S. A new approach to the interpretation of test scatter on the WISC-R. *Journal of Learning Disabilities*, 1976, *9*, 160-168. (a).

Kaufman, A. S. Do normal children have flat ability profiles? *Psychology In The Schools*, 1976, *13*, 284-285. (b)

Kaufman, A. S. Verbal-Performance IQ discrepancies in the WISC-R. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 739-744. (c)

Kaufman, A. S. *Intelligent testing with the WISC-R*. New York: John Wiley and Sons, 1979. (a)

Kaufman, A. S. WISC-R research: Implications for interpretation. *School Psychology Digest*, 1979, *8*, 5-27. (b)

Kaufman, A. S. Issues in psychological assessment: Interpreting the WISC-R intelligently. In B. B. Lahey & A. E. Kazdin (Eds.), *Advances in clinical child psychology* (Vol. 3). New York: Plenum Press, 1980.

Kaufman, A. S. The WISC-R and learning disabilities assessment: State of the art. *Journal of Learning Disabilities*, 1981, *14*, 520-526.

Kaufman, A. S. The impact of WISC-R research for school psychologists. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology*. New York: Wiley, 1982.

Kaufman, A. S., & Doppelt, J. E. Analysis of WISC-R standardization data in terms of the stratification variables. *Child Development*, 1976, *47*, 165-171.

Kelderman, H., Mellenbergh, G. J., & Elshout, J. J. Guilford's facet theory of intelligence: An empirical comparison of models. *Multivariate Behavioral Research*, 1981, *16*, 37-61.

King, A. R. *The school at Mopass: A problem of identity*. San Francisco: Holt, Rinehart & Winston, 1967.

Kleinfeld, J. S. *Achievement profiles of native ninth graders*. Fairbanks: University of Alaska, 1970. (ERIC Document Reproduction Service No. ED 045 282) (a)

Kleinfeld, J. S. *Cognitive strengths of Eskimos and implications for education*. Fairbanks: University of Alaska, 1970. (ERIC Document Reproduction Service No. ED 045 281) (b)

Kleinfeld, J. S. Sources of parental ambivalence toward education in an Aleut community. *Journal of American Indian Education*, 1971, *10*(2), 8-14. (a)

Kleinfeld, J. S. Visual memory in village Eskimo and urban Caucasian children. *Arctic*, 1971, *24*, 132-138. (b)

Kleinfeld, J. S. Classroom climate and the verbal participation of Indian and Eskimo students in integrated classrooms. *Journal of Educational Research*, 1973, *67*, 51-52. (a)

Kleinfeld J. S. Intellectual strengths in culturally different groups: An Eskimo illustration. *Review of Educational Research*, 1973, *43*, 341-359. (b)

Kleinfeld, J. S. Positive stereotyping: The cultural relativist in the classroom. *Human Organization*, 1975, *34*, 269-274.

Knowles, D. W., & Boersma, F. J. A comparison of optional shift performance and language skills in middle class and Canadian Indian children. *Canadian Journal of Behavioral Science*, 1971, *3*, 246-258.

Koppitz, E. M. *The Bender Gestalt Test for young children*. New York: Grune & Stratton, 1963.

Korth, B. A significance test for congruence coefficients for Cattell's factors matched by scanning. *Multivariate Behavioral Research*, 1978, *13*, 419-430.

Kroonenberg, P. M., & Lewis, C. Methodological issues in the search for a factor model: Exploration through confirmation. *Journal of Educational Statistics*, 1982, *7*, 69-89.

Laboratory of Comparative Human Cognition. What's cultural about cross-cultural cognitive psychology. *Annual Review of Psychology*, 1979, *30*, 145-172.

Lawley, D. N., & Maxwell, A.E. *Factor analysis as a statistical method* (2nd. ed.). London: Butterworths, 1971.

Leinert, G. A., & Croft, H. W. Studies on the factor structure of intelligence in children, adolescents, and adults. *Vita Humana*, 1964, *7*, 147-163.

Ling, D., McCoy, R. H., & Levison, E. D. The incidence of middle ear disease and its educational implications among Baffin Island Eskimo children. *Canadian Journal of Public Health*, 1969, *60*, 385-390.

Linn, R. L. A Monte Carlo approach to the number of factors problem. *Psychometrika*, 1968, *33*, 37-71.

Lomax, R. G. A guide to LISREL-type structural equation modeling. *Behavior Research Methods and Instrumentation*, 1982, *14*, 1-8.

Lowry, L. M. Differences in visual perception and auditory discrimination between American Indian and white kindergarten children. *Journal of Learning Disabilities*, 1970, *3*, 359-363.

Lutey, C. *Individual intelligence testing: A manual and sourcebook* (2nd. ed.). Greeley, Colo.: Carol L. Lutey, 1977.

Lyle, J. G., & Johnson, E. G. Analysis of WISC Coding: 3. Writing and copying speed, and motivation. *Perceptual and Motor Skills*, 1973, *36*, 211-214.

Lyle, J. G., & Johnson, E. G. Analysis of WISC Coding: 5. Prediction of Coding and performance. *Perceptual and Motor Skills*, 1974, *39*, 111-114.

MacArthur, R. S. *MacKenzie District norming project*, Report submitted to Chief, Education Division, Department of Northern Affairs and National Resources, 1965

MacArthur, R. S. Sex differences in field dependence for the Eskimo. *International Journal of Psychology*, 1967, *2*, 130-140.

MacArthur, R. S. Assessing the intellectual potential of native Canadian pupils: A summary. *Alberta Journal of Educational Research*, 1968, *14*, 115-122.

MacArthur, R. S. Some cognitive abilities of Eskimo, white, and Indian-Metis pupils aged 9 to 12 years. *Canadian Journal of Behavioral Science*, 1969, *1*, 50-59.

MacArthur, R. S. Some ability patterns: central Eskimos and Nsenga Africans. *International Journal of Psychology*, 1973, *8*, 239-247.

MacArthur, R. S. Differential ability patterns: Inuit, Nsenga, and Canadian whites. In J. W. Berry & W. J. Lonner (Eds.), *Applied cross-cultural psychology*. Amsterdam: Swets and Zeitlinger, 1975. (a)

MacArthur, R. S. Ecology, culture, and cognitive development: Canadian Native youth. in S. Schluderman, (Chm.), *Psychological development of the ethnic child and adolescent*. Symposium presented at a meeting of the Canadian Ethnic Studies Association, Winnipeg, Canada, 1975. (b)

Massey, F. J. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 1951, *46*, 68-78.

Matarazo, J. D. *Wechsler's measurement and appraisal of adult intelligence* (5th ed.). New York: Oxford University Press, 1972.

Matheson, D. W. Simultaneous-successive interpretation of the WISC-R: Making the most of indeterminacy. *Journal of Psychoeducational Assessment*, in press.

McDiarmid, G. L. *The hazards of testing Indian children*. np:nn, 1971. (ERIC Document Reproduction Service No. ED 055 692)

McDonald, R. P. & Mulaik, S. A. Determinacy of common factors: A nontechnical review. *Psychological Bulletin*, 1979, *86*, 297-306.

McFie, J. *Assessment of organic intellectual impairment*, London: Academic Press, 1975.

McGaw, B., & Jöreskog, K. G. Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status. *British Journal of Mathematical and Statistical Psychology*, 1971, *24*, 154-168.

McShane, D. A., & Plas, J. M. Wechsler scale performance patterns of American Indian children. *Psychology in the Schools*, 1982, *19*, 8-17.

Mercer, J. R. *SOMPA technical manual*. New York: The Psychological Corporation, 1979.

Mercer, J. R., & Ysseldyke, J. Designing diagnostic-intervention programs. In T. Oakland (Ed.), *Psychological and educational assessment of minority children*. New York: Brunner/Mazel, 1977.

Meredith, W. Rotation to achieve factorial invariance. *Psychometrika*, 1964, *29*, 187-206.

Messick, S. Beyond structure: In search of functional models of psychological process. *Psychometrika*, 1972, *37*, 357-375.

Messick, S. Test validity and the ethics of assessment. *American Psychologist*, 1980, *35*, 1012-1027.

Mickelson, N. I., & Galloway, C. G. Cumulative language deficit among Indian children. *Exceptional Children*, 1969, *36*, 187-190.

Mickelson, N. I., & Galloway, C. G. Verbal concepts of Indian and non-Indian school beginners. *Journal of Educational Research*, 1973, *67*, 55-56.

Mishra, S. P. Relationship of WISC-R factor scores to academic achievement and classroom behaviors of native American Navajos. *Measurement and Evaluation in Guidance*, 1981, *14*, 26-30.

Mishra, S. P., & Lord, J. Reliability and predictive validity of the WISC-R with Native-American Navajos. *Journal of School Psychology*, 1982, *20*, 150-154.

More, A. J., & Oldridge, B. An approach to non-discriminatory assessment of native Indian children. *B. C. Journal of Special Education*, 1980, *4*, 51-59.

Mueller, H. H., Dash, U. N., Matheson, D. W., & Short, R. H. WISC-R subtest patterning of below average, average, and above average IQ children: A meta-analysis. *Alberta Journal of Educational Research*, in press.

Mueller, H. H., Matheson, D. W., & Short, R. H. Bannatyne--recategorized WISC-R patterns of retarded, learning disabled, normal and intellectually superior children: A meta-analysis. *Mental Retardation/ Learning Disabilities Bulletin*, in press.

Mulaik, S. A. *The foundations of factor analysis*. New York: McGraw-Hill, 1972.

Mulcahy, R. F., & Watters, B. *Phase I: NWT norming project final report*. Yellowknife, NWT: Government of the Northwest Territories, 1982.

Naglieri, J. A. Does the WISC-R measure Verbal intelligence for nonEnglish-speaking children. *Psychology in the Schools*, 1982, *19*, 478-479.

Naglieri, J. A., Kamphaus, R. W., & Kaufman, A. S. The Luria-Das simultaneous-successive model applied to the WISC-R. *Journal of Psychoeducational Assessment*, 1983, *1*, 25-34.

Nesselroade, J. R., & Baltes, P. B. On a dilemma of comparitive factor analysis: A study of factor matching based on random data. *Educational and Psychological Measurement*, 1970, *30*, 935-948.

Nesselroade, J. R., Baltes, P. B., & Labouvie, E. W. Evaluating factor invariance in oblique space: Baseline data generated from random numbers. *Multivariate Behavioral Research*, 1971, *6*, 233-241.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. *Statistical package for the social sciences* (2nd ed.). New York: McGraw-Hill, 1975.

Nunnally, J. C. *Psychometric theory* (2nd ed.). New York: McGraw-Hill, 1978.

Oakland, T. An evaluation of the ABIC, pluralistic norms, and Estimated Learning Potential. *Journal of School Psychology*, 1980, *18*, 3-11.

Oakland, T., & Feigenbaum, D. Multiple sources of bias on the WISC-R and Bender-Gestalt Test. *Journal of Consulting and Clinical Psychology*, 1979, *47*, 968-974.

O'Connell, S. Television's impact on the Eskimo. *North Nord*, 1975, *22*(6), 34-37.

Olson, D. R., & MacArthur, R. S. The effect of foreign language background on intelligence test performance. *Alberta Journal of Educational Research*, 1962, *8*, 157-167.

Peck, R. L. A comparative analysis of the performance of Indian and white children from north central Montana on the Wechsler Intelligence Scale for Children (Doctoral dissertation, Montana State University, 1973). *Dissertation Abstracts International*, 1973, *33*, 4097A.

Pennell, R. Routinely computable confidence intervals for factor loadings using the jack-knife. *British Journal of Mathematical and Statistical Psychology*, 1972, *25*, 107-114.

Petersen, C. R., & Hart, D. H. Factor structure of the WISC-R for a clinic-referred population and specific subgroups. *Journal of Consulting and Clinical Psychology*, 1979, *47*, 643-645.

Phillips, D. C. What do the researcher and the practitioner have to offer each other. *Educational Researcher*, 1980, *11*, 17-20; 24.

Preston, C. E. Psychological testing with northwest coast Alaskan Eskimos. *Genetic Psychology Monographs*, 1964, *69*, 323-419.

Ramanaiah, N. V., O'Donnell, J. P., & Ribich, F. Multiple-group factor analysis of the Wechsler Intelligence Scale for Children. *Journal of Clinical Psychology*, 1976, *32*, 829-831.

Raskin, L. M., Bloom, A. S., Klee, S. H., & Reese, A. The assessment of developmentally disabled children with the WISC-R, Binet and other tests. *Journal of Clinical Psychology*, 1978, *34*, 111-114.

Rattan, M. S., & MacArthur, R. S. Longitudinal prediction of school achievement for Metis and Eskimo pupils. *Alberta Journal of Educational Research*, 1968, *14*, 37-41.

Raven, J. C. *Guide to the Standard Progressive Matrices*. London: H. K. Lewis, 1960.

Raven, J. C., Court, J. H., & Raven, J. *Manual for Raven's Progressive Matrices and vocabulary scales*. London: H. K. Lewis, 1977.

Reschly, D. WISC-R factor structure among Anglos, blacks, Chicanos, and native-American Papagos. *Journal of Consulting and Clinical Psychology*, 1978, *46*, 417-422.

Reschly, D. J. Nonbiased assessment. In G. D. Phye & D. J. Reschly (Eds.), *School psychology: Perspectives and issues*. New York: Academic Press, 1979.

Reschly, D. J., & Reschly, J. E. Validity of WISC-R factor scores in predicting achievement and attention for four sociocultural groups. *Journal of School Psychology*, 1979, *17*, 355-361.

Reschly, D. J., & Sabers, D.L. Analysis of test bias in four groups with the regression definition. *Journal of Educational Measurement*, 1979, *16*, 1-9.

Reynolds, C. R. Methods for detecting construct and predictive bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias*. Baltimore, Md.: John Hopkins University Press, 1982.

Reynolds, C. R., & Gutkin, T. B. Stability of the WISC-R factor structure across sex at two age levels. *Journal of Clinical Psychology*, 1980, *36*, 775-777.

Ricks, J. H. Local norms - when and why. *Test Service Bulletin, The Psychological Corporation*, 1981, *58*, 1-6.

Rohner, R. P. Factors influencing the academic performance of Kwakiutl children in Canada. *Comparative Educational Review*, 1965, *9*, 331-340.

Rosenbach, J. H., & Mowder, B. A. Test bias: The other side of the coin. *Psychology in the Schools*, 1981, *18*, 450-454.

Rugel, R. The factor structure of the WISC in two populations of disabled readers. *Journal of Learning Disabilities*, 1974, *7*, 581-585. (a)

Rugel, R. P. WISC subtest scores of disabled readers: a review with respect to Bannatyne's recategorization. *Journal of Learning Disabilities*, 1974, *7*, 57-64. (b)

Salvia, J., & Ysseldyke, J. E. *Assessment in special and remedial education* (2nd. ed.). Boston: Houghton Mifflin, 1981.

Sandoval, J. The WISC-R and internal evidence of test bias with minority groups. *Journal of Consulting and Clinical Psychology*, 1979, *47*, 919-927.

Sandoval, J. The WISC-R factorial validity for minority groups and Spearman's hypothesis. *Journal of School Psychology*, 1982, *20*, 198-204.

Satterly, D. J. Covariation of cognitive styles, intelligence, and achievement. *British Journal of Educational Psychology*, 1979, *49*, 179-181.

Sattler, J. M. *Assessment of children's intelligence and special abilities* (2nd. ed.). Boston: Allyn and Bacon, 1982.

Sattler, J. M., Andrea, J. R., Squire, L. S., Wisely, R., & Maloy, C. F. Examiner scoring of ambiguous WISC-R responses. *Psychology in the Schools*, 1978, *15*, 486-489.

Schaefer, O., Timmermans, J. F. W., Eaton, R. D. P., & Matthews, A. R. General and nutritional health in two Eskimo populations at different stages of acculturation. *Canadian Journal of Public Health*, 1980, *71*, 397-405.

Schönemann, P. H. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 1966, *31*, 1-10.

Schönemann, P. H. Factorial definitions of intelligence: Dubious legacy of dogma in data analysis. In I. Borg (Ed.). *Multidimensional data representations: When and why*. Ann Arbor: Mathesis, 1981.

Schooler, D. L., Beebe, M. C., & Hoepke, T. Factor analysis of WISC-R scores for children identified as learning disabled, educable mentally impaired, and emotionally impaired. *Psychology in the Schools*, 1978, *15*, 478-485.

Schubert, J., & Cropley, A. J. Verbal regulation of behavior and I.Q. in Canadian Indian and white children. *Developmental Psychology*, 1972, 7, 295-301.

Scollon, R. *The context of the informant narrative performance: From sociolinguistics to ethnolinguistics at Fort Chipewyan, Alberta*. Ottawa: National Museums of Canada, 1979.

Serpell, R. *Selective attention and the interference between first and second languages*. Lusaka, Zambia: Institute for Social Research, University of Zambia, 1968.

Seyfort, B., Spreen, O., & Lahmer, V. A. A critical look at the WISC-R with native Indian children. *Alberta Journal of Educational Research*, 1980, *26*, 14-29.

Shiek, D. A., & Miller, J.E. Validity generalization of the WISC-R factor structure with 10 1/2-year-old children. *Journal of Consulting and Clinical Psychology*, 1978, *46*, 583.

Silverstein, A. B. Factor structure of the Wechsler Intelligence Scale for Children for three ethnic groups.

Journal of Educational Psychology, 1973, *65*, 408-410.

Silverstein, A. B. Variance components in the subtests of the WISC-R. *Psychological Reports*, 1976, *39*, 1109-1110.

Silverstein, A. B. Alternative factor analytic solutions for the Wechsler Intelligence Scale for Children - Revised. *Educational and Psychological Measurement*, 1977, *37*, 121-124.

Silverstein, A. B. Cluster analysis of the Wechsler Intelligence Scale for Children - Revised. *Educational and Psychological Measurement*, 1980, *40*, 51-54.

Silverstein, A. B. Reliability and abnormality of test score differences. *Journal of Clinical Psychology*, 1981, *37*, 392-394.

Silverstein, A. B. Alternative multiple-group solutions for the WISC and WISC-R. *Journal of Clinical Psychology*, 1982, *38*, 166-168.

Silverstein, A. B., & Legutki, G. Direct comparisons of the factor structures of the WISC and the WISC-R. *Psychology in the Schools*, 1982, *19*, 5-7.

Skakun, E. N. *A Monte Carlo approach to the factorial invariance problem using the orthogonal Procrustes solution*. Unpublished M.Ed. thesis, University of Alberta, 1971.

Smirnov, N. M. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 1948, *19*, 279-281.

Smith, C. A. & Lawley, D. N. *Mental testing of the Hebridean children in Gaelic and English*. London: University of London Press, 1948.

Smith, M. D. Stability of WISC-R subtest profiles for learning disabled children. *Psychology in the Schools*, 1978, *15*, 4-7.

Smith, M. D., Coleman, J. M., Dockecki, P. R., & Davis, E. E. Intellectual characteristics of school labeled learning disabled children. *Exceptional Children*, 1977, *43*, 352-357. (a)

Smith, M. D., Coleman, J. M., Dockecki, P. R., & Davis, E. E. Recategorized WISC-R scores of learning disabled children, *Journal of Learning Disabilities*, 1977, *10*, 437-443. (b)

Sobotka, K. R., & Black, F. W. A procedure for the rapid computation of WISC-R factor scores. *Journal of Clinical Psychology*, 1978, *34*, 117-119.

Sörbom, D. Detection of correlated errors in longitudinal data. *British Journal of Mathematical and Statistical Psychology*, 1975, *28*, 138-151.

Soveran, M. *From Cree to English, Part I: The sound system. Saskatoon, Sk: Indian and Northern Curriculum Resources Centre, University of Saskatchewan*, n.d.

Spearman, C. E. 'General intelligence' objectively determined and measured. *American Journal of Psychology*, 1904, *15*, 206-293.

Spearman, C. E. *The abilities of man: Their nature and measurement*. London: MacMillan, 1927.

Stedman, J. M., Lawlis, G., Cortner, R. H., & Achterberg, G. Relationships between WISC-R factors, Wide-Range Achievement Test scores, and visual-motor maturation in children referred for psychological evaluation. *Journal of Consulting and Clinical Psychology*, 1978, *46*, 869-872.

Sternberg, R. J. Factor theories of intelligence are all right almost. *Educational Researcher*, 1980, *9*(8), 6-13; 18.

Sternberg, R. J. Nothing fails like success: The search for an intelligent paradigm for studying intelligence. *Journal of Educational Psychology*, 1981, *73*, 142-155.

Sternberg, R. J. Reasoning, problem solving and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence*. New York: Cambridge University Press, 1982.

St. John, J., & Krichev, A. Northwestern Ontario Indian children and the WISC. *Psychology in the Schools*, 1976, *13*, 407-411.

Swanson, H. L., & Watson, B. L. *Educational and psychological assessment of exceptional children: Theories, strategies, and applications*. St. Louis: Mosby, 1982.

Swerdlik, M. E., & Schweitzer, J. A comparison of factor structures of the WISC and WISC-R. *Psychology in the Schools*, 1978, *15*, 166-172.

Swyter, L. J., & Michael, W. B. The interrelationships of four measures hypothesized to represent field dependence - field independence construct. *Educational and*

Psychological Measurement, 1982, *42*, 877-888.

Taylor, L. J., & Skanes, G. Psycholinguistic abilities of children in isolated communities of Labrador. *Canadian Journal of Behavioral Science*, 1975, *7*, 30-39.

Taylor, L. J., & Skanes, G. R. Cognitive abilities in Inuit and White children from similar environments. *Canadian Journal of Behavioral Science*, 1976, *8*, 1-8. (a)

Taylor, L. J., & Skanes, G. R. Level I and level II intelligence in Inuit and White children from similar environments. *Journal of Cross-Cultural Psychology*, 1976, *7*, 157-168. (b)

Teeter, A., Moore, C., & Petersen, J. D. WISC-R verbal and performance abilities of native American students referred for school learning problems. *Psychology in the Schools*, 1982, *19*, 39-44.

Tellegen, A., & Briggs, P. F. Old wine in new skins: Grouping Wechsler subtests into new scales. *Journal of Consulting Psychology*, 1967, *31*, 499-506.

Thurstone, L. L. *Primary mental abilities*. Chicago: University of Chicago Press, 1938.

Toukamaa, P., & Skutnabb-Kangas, T. *The intensive teaching of the mother tongue to migrant children at pre-school age*. Tampere: University of Tampere, Finland, 1977.

Tucker, J. A. Operationalizing the diagnostic-intervention process. In T. Oakland (Ed.), *Psychological and educational assessment of minority children*. New York: Brunner/Mazel, 1977.

Tucker, L. R., & Lewis, C. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 1973, *38*, 1-10.

Vallee, F. G. Differentiation among the Eskimos in some Canadian arctic settlements. In V. F. Valentine and F. G. Vallee (Eds.), *Eskimo of the Canadian Arctic*, Toronto: McClelland & Stewart, 1968.

Vance, H. B., & Engin, A. Analysis of cognitive abilities of black children's performances on the WISC-R. *Journal of Clinical Psychology*, 1978, *34*, 452-456.

Vance, H. B., & Gaynor, P. E. A note on cultural difference as reflected in the Wechsler Intelligence Scale for Children. *Journal of Genetic Psychology*, 1976, *129*, 171-172.

Vance, H. B., Singer, M. G., Kitson, D. L., & Brenner, O. C. WISC-R profile analysis in differentiating LD from ED children. *Journal of Clinical Psychology*, 1983, *39*, 125-132.

Vance, H. B., Wallbrown, F. H., & Fremont, T. S. The abilities of retarded students: Further evidence regarding the stimulus trace factor. *Journal of Psychology*, 1978, *100*, 77-82.

Vandenberg, S. G., & Hakstian, A. R. Cultural influences on cognition: a reanalysis of Vernon's data. *International Journal of Psychology*, 1978, *13*, 251-279.

Van Hagen, J., & Kaufman, A.S. Factor analysis of the WISC-R for a group of mentally retarded children and adolescents. *Journal of Consulting and Clinical Psychology*, 1975, *43*, 661-667.

Velicer, W. F., Peacock, A. C., & Jackson, D. N. A comparison of component and factor patterns: a Monte Carlo approach. *Multivariate Behavioral Research*, 1982, *17*, 371-388.

Vernon, P. E. Ability factors and environmental influences. *American Psychologist*, 1965, *20*, 723-733.

Vernon, P. E. Educational and intellectual development among Canadian Indians and Eskimos: Part I. *Educational Review*, 1966, *18*, 79-91. (a)

Vernon, P. E. Educational and intellectual development among Canadian Indians and Eskimos: Part II. *Educational Review*, 1966, *18*, 186-195. (b)

Vernon, P. E. *Intelligence and cultural environment*. London: Methuen, 1969.

Vernon, P. E. The distinctiveness of field independence. *Journal of Personality*, 1972, *40*, 366-391.

Wachtel, P. L. Style and capacity in analytic functioning. *Journal of Personality*, 1968, *36*, 202-212.

Wachtel, P. L. Field dependence and psychological differentiation: Reexamination. *Perceptual and Motor Skills*, 1972, *35*, 179-189.

Wallbrown, F. H., Blaha, J., Wallbrown, J. D., & Engin, A. W. The hierarchical factor structure of Wechsler Intelligence Scale for Children - Revised. *Journal of Psychology*, 1975, *89*, 223-235.

Watson, L. Television and its early social effect among

Rankin Inlet Inuit. *Musk-ox*, 1980, *27*, 60-66.

Watters, B. Special education in the Northwest Territories. In M. Csapo & L. Goguen (Eds.), *Special education across Canada*. Vancouver: Centre for Human Development and Research, 1980.

Watters, B. Special education in the Northwest Territories: A review. *Canadian Journal of Exceptionality*, in press.

Wattie, D. K. F. Education in the Canadian arctic. *Polar Record*, 1968, 293-304.

Wechsler, D. *Manual for the Wechsler Intelligence Scale for Children*. New York: Psychological Corporation, 1949.

Wechsler, D. *Manual for the Wechsler Adult Intelligence Scale*. New York: Psychological Corporation, 1955.

Wechsler, D. *Manual for the Wechsler Preschool and Primary Scale of Intelligence*. New York: Psychological Corporation, 1967.

Wechsler, D. *Manual for the Wechsler Intelligence Scale for Children - Revised*. New York: Psychological Corporation, 1974.

Weinberg, J., Diller, L., Gerstman, L., & Schulman, P. Digit Span in right and left hemiplegics. *Journal of Clinical Psychology*, 1972, *28*, 361.

Werts, C. E., Rock, D. A., Linn, R. L., & Jöreskog, K. G. Comparison of correlations, variances, covariances, and regression weights with and without measurement error. *Psychological Bulletin*, 1976, *83*, 1007-1013.

West, L. W., & MacArthur, R. S. An evaluation of selected intelligence tests for two samples of Metis and Indian children. *Alberta Journal of Educational Research*, 1964, *10*, 17-27.

Wikoff, R. L. The WISC-R as a predictor of achievement. *Psychology in the Schools*, 1979, *16*, 364-366.

Williams, R. L. Abuses and misuses in testing black children. *Counseling Psychologist*, 1971, *2*, 62-77.

Wiltshire, E. B., & Gray, J. E. Draw-a-Man and Raven's Progressive Matrices (1938) intelligence test performance of reserve Indian children. *Canadian Journal of Behavioral Science*, 1969, *1*, 119-122.

Witkin, H. A. Cognitive styles across cultures. In J. W. Berry & P. R. Dasen (Eds.), *Culture and cognition:*

Readings in cross-cultural psychology. London: Methuen, 1974.

Witkin, H. A., Moore, C. A., Goodenough, D. R., & Cox, P. W. Field-dependent and field-independent cognitive styles and their educational implications. *Review of Educational Research*, 1977, *47*, 1-64.

Wrigley, C. S., & Neuhaus, J. O. The matching of two sets of factors. *American Psychologist*, 1955, *10*, 418-419.

Ysseldyke, J. E., & Mirken, P. K. The use of assessment information to plan interventions: A review of the research. In C. R. Reynolds & T. B. Gutkin (Eds.), *Handbook of School Psychology*, New York: Wiley, 1982.

Zarske, J. A., & Moore, C. L. Recategorized WISC-R scores for non-handicapped, learning disabled, educationally disadvantaged and regular classroom Navajo children. *School Psychology Review*, 1982, *11*, 319-323. (a)

Zarske, J. A., & Moore, C. L. Recategorized WISC-R scores of learning disabled Navajo Indian children. *Psychology in the Schools*, 1982, *19*, 156-159. (b)

Zarske, J. A., Moore, C. L., & Petersen, J. D. WISC-R factor structure for diagnosed learning disabled Navajo and Papago children. *Psychology in the Schools*, 1981, *18*, 402-407.

Zarske, J. A., Moore, C. L., & Petersen, J. D. Tripping over the leaves: A response to Naglieri. *Psychology in the Schools*, 1982, *19*, 480-481.

Zingale, S. A., & Smith, M. D. WISC-R patterns for learning disabled children at three SES levels. *Psychology in the Schools*, 1978, *15*, 199-204.

Zinkas, P. W., Gottlieb, M., & Schapiro, M. Developmental and psychoeducational sequelae of chronic otitis media. *American Journal of Diseases of Children*, 1978, *132*, 1100-1104.